

taller-3

April 22, 2024

- **Integrante 1:** Lina María Gómez Mesa
 - **Integrante 2:** María Catalina Ibáñez Piñeres
-

1 Contexto y objetivos.

En la actualidad, la cantidad de artículos publicados en Internet está generando una gran ola de información accesible por cualquier usuario, dando a conocer diferentes puntos de vista, opiniones, información e investigaciones sobre diferentes temas de interés.

Esta gran cantidad de información no solo permite una búsqueda exhaustiva sobre un tema, también permite realizar un análisis sobre la tendencia de los diferentes temas que estén dando de qué hablar en una sociedad. Es por ello que un grupo de expertos se ha dado la tarea de analizar 10.000 artículos web y clasificarlos para poder establecer un análisis de los temas en la actualidad.

Para ello, como experto en análisis con machine learning, le han pedido que construya un modelo capaz de clasificar los nuevos artículos, realice un análisis de cuáles son los temas que dan de qué hablar y automatice el proceso de selección y búsqueda de diferentes artículos.

Objetivos de desarrollo:

- Realizar el análisis y limpieza de textos.
- Explorar las diferentes técnicas de transformación de datos no estructurados.
- Establecer el mejor modelo basado en una red neuronal.

Datos: La fuente de los datos la puedes encontrar en [News Articles Classification Dataset for NLP & ML](#).

2 Entendimiento del negocio.

Para tener un mejor detalle sobre el comportamiento de las variables, solicitamos a la organización el diccionario de datos y nos suministró la siguiente información:

ATRIBUTO	DEFINICIÓN
headlines	Titular del artículo.
description	Reseña del artículo.
content	Contenido del artículo.

ATRIBUTO	DEFINICIÓN
url	Dirección web del artículo.
category	Representa la temática del artículo.

3 Actividades a realizar.

1. Realizar el análisis exploratorio de componentes principales en la información.
2. Identificar el número de componentes principales apropiado el procesamiento. Genera una tabla comparativa y los gráficos que apoyen este proceso. Recuerda que no deben truncarse los textos. Por último, la elección del número de componentes debe estar debidamente justificada.
3. Construir la red neuronal tomando como insumo los componentes principales procesados en el punto anterior.
4. Construir las gráficas de entrenamiento, validación. Debes interpretar los resultados obtenidos para este modelo base.
5. Realizar la identificación de hiperparámetros, justificando la elección de los valores correspondientes.

NOTA: La calificación será sobre notebook ejecutado y cargado en Bloque Neón junto con el archivo HTML.

4 0. Importar librerías

```
[1]: #Manejo de datos
import pandas as pd
import numpy as np
import scipy

#Visualización de datos
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

#Analisis profundo de datos

from ydata_profiling import ProfileReport

#Entrenamiento del modelo
import sklearn
from sklearn.decomposition import PCA, TruncatedSVD
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.model_selection import train_test_split
#from sklearn.metrics import mean_squared_error, r2_score
#from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import LabelEncoder
```

```

#from sklearn.compose import ColumnTransformer, make_column_selector
from sklearn.metrics import classification_report, confusion_matrix, PrecisionRecallDisplay

#Textos

import contractions
import inflect
import nltk
import re, string, unicodedata
from nltk import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer, WordNetLemmatizer
from polyglot.detect import Detector
from wordcloud import WordCloud, STOPWORDS

#Tensorflow y keras
import tensorflow as tf
from keras.callbacks import EarlyStopping
from keras.models import Sequential
from keras.layers import Dense
from keras.utils import plot_model

#Sistema operativo
import os
import os.path as osp

#Librerías extras
import itertools
from datetime import datetime

print(f"La versión de sklearn es: {sklearn.__version__}")
print(f'La versión de Tensor Flow es:', tf.__version__)

```

```

/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-packages/tqdm/auto.py:21:
TqdmWarning: IPProgress not found. Please update jupyter and ipywidgets. See
https://ipywidgets.readthedocs.io/en/stable/user_install.html
    from .autonotebook import tqdm as notebook_tqdm

```

```

La versión de sklearn es: 1.4.2
La versión de Tensor Flow es: 2.16.1

```

```
[ ]: nltk.download('all')
```

```

[nltk_data]  Downloading collection 'all'
[nltk_data]      |
[nltk_data]      | Downloading package abc to

```

```
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package abc is already up-to-date!
[nlk_data] | Downloading package alpino to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package alpino is already up-to-date!
[nlk_data] | Downloading package averaged_perceptron_tagger to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package averaged_perceptron_tagger is already up-
[nlk_data] |   to-date!
[nlk_data] | Downloading package averaged_perceptron_tagger_ru to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package averaged_perceptron_tagger_ru is already
[nlk_data] |   up-to-date!
[nlk_data] | Downloading package basque_grammars to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package basque_grammars is already up-to-date!
[nlk_data] | Downloading package bcp47 to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package bcp47 is already up-to-date!
[nlk_data] | Downloading package biocreative_ppi to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package biocreative_ppi is already up-to-date!
[nlk_data] | Downloading package blip_wsj_no_aux to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package blip_wsj_no_aux is already up-to-date!
[nlk_data] | Downloading package book_grammars to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package book_grammars is already up-to-date!
[nlk_data] | Downloading package brown to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package brown is already up-to-date!
[nlk_data] | Downloading package brown_tei to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package brown_tei is already up-to-date!
[nlk_data] | Downloading package cess_cat to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package cess_cat is already up-to-date!
[nlk_data] | Downloading package cess_esp to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package cess_esp is already up-to-date!
[nlk_data] | Downloading package chat80 to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package chat80 is already up-to-date!
[nlk_data] | Downloading package city_database to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package city_database is already up-to-date!
[nlk_data] | Downloading package cmudict to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
```

```
[nltk_data] |   Package cmudict is already up-to-date!
[nltk_data] | Downloading package comparative_sentences to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package comparative_sentences is already up-to-
[nltk_data] |   date!
[nltk_data] | Downloading package comtrans to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package comtrans is already up-to-date!
[nltk_data] | Downloading package conll2000 to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package conll2000 is already up-to-date!
[nltk_data] | Downloading package conll2002 to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package conll2002 is already up-to-date!
[nltk_data] | Downloading package conll2007 to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package conll2007 is already up-to-date!
[nltk_data] | Downloading package crubadan to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package crubadan is already up-to-date!
[nltk_data] | Downloading package dependency_treebank to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package dependency_treebank is already up-to-date!
[nltk_data] | Downloading package dolch to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package dolch is already up-to-date!
[nltk_data] | Downloading package europarl_raw to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package europarl_raw is already up-to-date!
[nltk_data] | Downloading package extended_omw to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package extended_omw is already up-to-date!
[nltk_data] | Downloading package floresta to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package floresta is already up-to-date!
[nltk_data] | Downloading package framenet_v15 to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package framenet_v15 is already up-to-date!
[nltk_data] | Downloading package framenet_v17 to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package framenet_v17 is already up-to-date!
[nltk_data] | Downloading package gazetteers to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package gazetteers is already up-to-date!
[nltk_data] | Downloading package genesis to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package genesis is already up-to-date!
[nltk_data] | Downloading package gutenberg to
```

```
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package gutenberg is already up-to-date!
[nlk_data] | Downloading package ieer to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package ieer is already up-to-date!
[nlk_data] | Downloading package inaugural to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package inaugural is already up-to-date!
[nlk_data] | Downloading package indian to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package indian is already up-to-date!
[nlk_data] | Downloading package jeita to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package jeita is already up-to-date!
[nlk_data] | Downloading package kimmo to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package kimmo is already up-to-date!
[nlk_data] | Downloading package knbc to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package knbc is already up-to-date!
[nlk_data] | Downloading package large_grammars to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package large_grammars is already up-to-date!
[nlk_data] | Downloading package lin_thesaurus to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package lin_thesaurus is already up-to-date!
[nlk_data] | Downloading package mac_morpho to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package mac_morpho is already up-to-date!
[nlk_data] | Downloading package machado to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package machado is already up-to-date!
[nlk_data] | Downloading package masc_tagged to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package masc_tagged is already up-to-date!
[nlk_data] | Downloading package maxent_ne_chunker to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package maxent_ne_chunker is already up-to-date!
[nlk_data] | Downloading package maxent_treebank_pos_tagger to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package maxent_treebank_pos_tagger is already up-
[nlk_data] |       to-date!
[nlk_data] | Downloading package moses_sample to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package moses_sample is already up-to-date!
[nlk_data] | Downloading package movie_reviews to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] |   Package movie_reviews is already up-to-date!
```

```
[nlk_data] | Downloading package mte_teip5 to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package mte_teip5 is already up-to-date!
[nltk_data] | Downloading package mwa_ppdb to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package mwa_ppdb is already up-to-date!
[nltk_data] | Downloading package names to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package names is already up-to-date!
[nltk_data] | Downloading package nombank.1.0 to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package nombank.1.0 is already up-to-date!
[nltk_data] | Downloading package nonbreaking_prefixes to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package nonbreaking_prefixes is already up-to-date!
[nltk_data] | Downloading package nps_chat to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package nps_chat is already up-to-date!
[nltk_data] | Downloading package omw to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package omw is already up-to-date!
[nltk_data] | Downloading package omw-1.4 to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package omw-1.4 is already up-to-date!
[nltk_data] | Downloading package opinion_lexicon to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package opinion_lexicon is already up-to-date!
[nltk_data] | Downloading package panlex_swadesh to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package panlex_swadesh is already up-to-date!
[nltk_data] | Downloading package paradigms to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package paradigms is already up-to-date!
[nltk_data] | Downloading package pe08 to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package pe08 is already up-to-date!
[nltk_data] | Downloading package perluniprops to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package perluniprops is already up-to-date!
[nltk_data] | Downloading package pil to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package pil is already up-to-date!
[nltk_data] | Downloading package pl196x to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package pl196x is already up-to-date!
[nltk_data] | Downloading package porter_test to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package porter_test is already up-to-date!
```

```
[nlk_data] | Downloading package ppattach to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package ppattach is already up-to-date!
[nlk_data] | Downloading package problem_reports to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package problem_reports is already up-to-date!
[nlk_data] | Downloading package product_reviews_1 to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package product_reviews_1 is already up-to-date!
[nlk_data] | Downloading package product_reviews_2 to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package product_reviews_2 is already up-to-date!
[nlk_data] | Downloading package propbank to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package propbank is already up-to-date!
[nlk_data] | Downloading package pros_cons to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package pros_cons is already up-to-date!
[nlk_data] | Downloading package ptb to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package ptb is already up-to-date!
[nlk_data] | Downloading package punkt to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package punkt is already up-to-date!
[nlk_data] | Downloading package qc to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package qc is already up-to-date!
[nlk_data] | Downloading package reuters to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package reuters is already up-to-date!
[nlk_data] | Downloading package rslp to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package rslp is already up-to-date!
[nlk_data] | Downloading package rte to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package rte is already up-to-date!
[nlk_data] | Downloading package sample_grammars to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package sample_grammars is already up-to-date!
[nlk_data] | Downloading package semcor to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package semcor is already up-to-date!
[nlk_data] | Downloading package senseval to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package senseval is already up-to-date!
[nlk_data] | Downloading package sentence_polarity to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package sentence_polarity is already up-to-date!
```

```
[nlk_data] | Downloading package sentiwordnet to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package sentiwordnet is already up-to-date!
[nlk_data] | Downloading package shakespeare to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package shakespeare is already up-to-date!
[nlk_data] | Downloading package sinica_treebank to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package sinica_treebank is already up-to-date!
[nlk_data] | Downloading package smultron to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package smultron is already up-to-date!
[nlk_data] | Downloading package snowball_data to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package snowball_data is already up-to-date!
[nlk_data] | Downloading package spanish_grammars to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package spanish_grammars is already up-to-date!
[nlk_data] | Downloading package state_union to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package state_union is already up-to-date!
[nlk_data] | Downloading package stopwords to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package stopwords is already up-to-date!
[nlk_data] | Downloading package subjectivity to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package subjectivity is already up-to-date!
[nlk_data] | Downloading package swadesh to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package swadesh is already up-to-date!
[nlk_data] | Downloading package switchboard to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package switchboard is already up-to-date!
[nlk_data] | Downloading package tagsets to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package tagsets is already up-to-date!
[nlk_data] | Downloading package timit to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package timit is already up-to-date!
[nlk_data] | Downloading package toolbox to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package toolbox is already up-to-date!
[nlk_data] | Downloading package treebank to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package treebank is already up-to-date!
[nlk_data] | Downloading package twitter_samples to
[nlk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nlk_data] | Package twitter_samples is already up-to-date!
```

```
[nltk_data] | Downloading package udhr to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] | Package udhr is already up-to-date!
[nltk_data] | Downloading package udhr2 to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] | Package udhr2 is already up-to-date!
[nltk_data] | Downloading package unicode_samples to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] | Package unicode_samples is already up-to-date!
[nltk_data] | Downloading package universal_tagset to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] | Package universal_tagset is already up-to-date!
[nltk_data] | Downloading package universal_treebanks_v20 to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] | Package universal_treebanks_v20 is already up-to-
[nltk_data] |   date!
[nltk_data] | Downloading package vader_lexicon to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] | Package vader_lexicon is already up-to-date!
[nltk_data] | Downloading package verbnet to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] | Package verbnet is already up-to-date!
[nltk_data] | Downloading package verbnet3 to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] | Package verbnet3 is already up-to-date!
[nltk_data] | Downloading package webtext to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] | Package webtext is already up-to-date!
[nltk_data] | Downloading package wmt15_eval to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] | Package wmt15_eval is already up-to-date!
[nltk_data] | Downloading package word2vec_sample to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] | Package word2vec_sample is already up-to-date!
[nltk_data] | Downloading package wordnet to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] | Package wordnet is already up-to-date!
[nltk_data] | Downloading package wordnet2021 to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] | Package wordnet2021 is already up-to-date!
[nltk_data] | Downloading package wordnet2022 to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] | Package wordnet2022 is already up-to-date!
[nltk_data] | Downloading package wordnet31 to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] | Package wordnet31 is already up-to-date!
[nltk_data] | Downloading package wordnet_ic to
[nltk_data] |   /Users/mariacatalinaibanezpineres/nltk_data...
```

```
[nltk_data] |   Package wordnet_ic is already up-to-date!
[nltk_data] | Downloading package words to
[nltk_data] |     /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package words is already up-to-date!
[nltk_data] | Downloading package ycoe to
[nltk_data] |     /Users/mariacatalinaibanezpineres/nltk_data...
[nltk_data] |   Package ycoe is already up-to-date!
[nltk_data] |
[nltk_data] Done downloading collection all
```

[]: True

5 1. Introducción a los datos

5.1 1.1. Configuración del entorno de kaggle

Se hace la conexión con kaggle para poder descargar la base de datos.

```
[3]: !ls -lha kaggle.json
! mkdir ~/.kaggle
! cp kaggle.json ~/.kaggle/
! chmod 600 ~/.kaggle/kaggle.json
```

```
-rw-r--r--@ 1 mariacatalinaibanezpineres  staff    75B Apr 12 09:57 kaggle.json
mkdir: /Users/mariacatalinaibanezpineres/.kaggle: File exists
```

Se verifica la conectividad con el entorno de kaggle.

```
[4]: !kaggle datasets list
```

ref	title			
size	lastUpdated	downloadCount	voteCount	usabilityRating
sudarshan24byte/online-food-dataset	Online Food			
Dataset	3KB	2024-03-02 18:50:30	25567	
502 0.9411765				
nbroad/gemma-rewrite-nbroad	gemma-rewrite-			
nbroad	8MB	2024-03-03 04:52:39	1610	
101 1.0				
sukhmandeepsinghbrar/most-subscribed-youtube-channel	Most Subscribed			
YouTube Channel	1KB	2024-04-10 20:33:05	928	
31 1.0				
sanyamgoyal401/customer-purchases-behaviour-dataset	Customer Purchases			
Behaviour Dataset	1MB	2024-04-06 18:42:01	1487	
37 1.0				
divu2001/restaurant-order-data	Restaurant Order			
Data	426KB	2024-04-10 08:33:29	597	

```

23 1.0
startalks/pii-models                                pii-models
1GB 2024-03-21 21:23:40      109      19 1.0
bhavikjikadara/student-study-performance          Student Study
Performance                                         9KB 2024-03-07 06:14:09      12506
160 1.0
fatemehmehrparvar/obesity-levels                  Obesity Levels
58KB 2024-04-07 16:28:30      1570      43 0.88235295
sahirmaharajj/employee-salaries-analysis          Employee Salaries
Analysis                                           101KB 2024-03-31 16:32:47      1360
41 1.0
mohdshahnawazaadil/credit-card-dataset          Credit Card
Dataset                                         66MB 2024-04-02 00:04:05      904
25 1.0
sahilnbajaj/loans-data                            Loans Data
213KB 2024-04-07 15:08:37      1074      29 1.0
sukhmandeepsinghbrar/housing-price-dataset        Housing Price
Dataset                                         780KB 2024-04-04 19:45:43      1667
31 1.0
sumyajitjalua/crop-datasets-for-all-indian-states-2010-2017 Crop Datasets for
All Indian States: 2010-2017 305KB 2024-04-09 15:16:46      453
24 1.0
willianoliveiragibin/time-the-internet           time the Internet
43KB 2024-03-28 18:36:19      969      29 1.0
sahirmaharajj/air-pollution-dataset              Air Pollution
Dataset                                         213KB 2024-04-07 13:14:48      1190
36 1.0
tanmay43sharma/goodreads-popular-books-dataset Popular Books
Dataset                                         130KB 2024-03-28 09:44:18      1356
29 0.9411765
joshuanaude/effects-of-alcohol-on-student-performance Effects of Alcohol
on Student Performance. 9KB 2024-03-25 12:08:03      1447
30 1.0
sahirmaharajj/electric-vehicle-population-size-2024 Electric Vehicle
Population by Country (2024) 275KB 2024-03-30 19:16:06      2215
56 1.0
sahirmaharajj/country-health-trends-dataset       Country Health
Trends Dataset                                     4KB 2024-04-10 10:57:26      322
26 1.0
jatinthakur706/most-watched-netflix-original-shows-tv-time Most watched
Netflix original shows (TV Time) 2KB 2024-03-27 09:01:21      2921
49 1.0

```

Se descarga la base de datos.

```
[5]: !kaggle datasets download banuprakashv/
↳news-articles-classification-dataset-for-nlp-and-ml
```

```
Dataset URL: https://www.kaggle.com/datasets/banuprakashv/news-articles-classification-dataset-for-nlp-and-ml  
License(s): Apache 2.0  
news-articles-classification-dataset-for-nlp-and-ml.zip: Skipping, found more recently modified local copy (use --force to force download)
```

```
[2]: ROOT_DIR = 'content'  
DATASET_NAME = 'news-articles-classification-dataset-for-nlp-and-ml'
```

```
[7]: print(f"!unzip {DATASET_NAME}.zip -d {ROOT_DIR}/{DATASET_NAME}")
```

```
!unzip news-articles-classification-dataset-for-nlp-and-ml.zip -d content/news-articles-classification-dataset-for-nlp-and-ml
```

Se descomprime el archivo en una carpeta previamente creada llamada content

```
[8]: #%%cd {ROOT_DIR}  
!mkdir content  
!mkdir content/{DATASET_NAME}  
!unzip {DATASET_NAME}.zip -d {ROOT_DIR}/{DATASET_NAME}
```

```
mkdir: content: File exists  
mkdir: content/news-articles-classification-dataset-for-nlp-and-ml: File exists  
Archive: news-articles-classification-dataset-for-nlp-and-ml.zip  
replace content/news-articles-classification-dataset-for-nlp-and-ml/business_data.csv? [y]es, [n]o, [A]ll, [N]one, [r]ename: ^C
```

Se genera la ruta del directorio para cargar la información.

```
[3]: DATA_DIR = f"{ROOT_DIR}/{DATASET_NAME}"  
print(DATA_DIR)
```

```
content/news-articles-classification-dataset-for-nlp-and-ml
```

5.2 1.2. Separación de la información.

Se listan los archivos dentro de la carpeta

```
[4]: csv_files = os.listdir(DATA_DIR)  
  
train_df = pd.DataFrame()  
test_df = pd.DataFrame()  
  
for csv_file in csv_files:  
    new_df = pd.read_csv(os.path.join(DATA_DIR, csv_file))  
    train, test = train_test_split(new_df, test_size=0.2, random_state=19)  
    train_df = pd.concat([train_df, train])  
    test_df = pd.concat([test_df, test])  
  
train_df.sample(5)
```

```
[4]:                                                 headlines \
556 Indian Open: Playing for his late father, Anga...
316 EPFO unveils FAQs on issues linked to higher p...
1102 First pictures of BTS members Jungkook and Jim...
1058 Amazon cuts jobs in music streaming unit
676 'Will be Leo's decision': Coach Scaloni hints ...

                                                 description \
556 The 33-year-old shot an impressive 71 to finis...
316 For those who will retire in the future, say i...
1102 Jimin and Jungkook started their mandatory mil...
1058 The cuts come even as Amazon reported third-qu...
676 Lionel Messi will be 39 for the 2026 World Cup...

                                                 content \
556 After all of the promising starts made by a st...
316 The Employees' Provident Fund Organisation (EP...
1102 All seven members of the K-pop group BTS have ...
1058 Amazon.com has begun cutting jobs in its Music...
676 Argentina's World Cup-winning coach Lionel Sca...

                                                 url          category
556 https://indianexpress.com/article/sports/golf/...      sports
316 https://indianexpress.com/article/business/epf...      business
1102 https://indianexpress.com/article/entertainmen...   entertainment
1058 https://indianexpress.com/article/technology/t...  technology
676 https://indianexpress.com/article/sports/footb...      sports
```

Se mira el número de instancias para cada uno de los conjuntos de datos.

```
[5]: train_count = train_df.shape[0]
test_count = test_df.shape[0]

print("-----SEPARACIÓN DE LA INFORMACIÓN-----")
print(f"--> Train: {train_count},")
print(f"--> Test: {test_count},")
```

-----SEPARACIÓN DE LA INFORMACIÓN-----

-> Train: 8,000
-> Test: 2,000

Se verifican las categorías

```
[6]: train_df["category"].value_counts()
```

```
[6]: category
entertainment    1600
education        1600
```

```
business          1600
technology        1600
sports            1600
Name: count, dtype: int64
```

```
[7]: test_df["category"].value_counts()
```

```
[7]: category
entertainment    400
education        400
business         400
technology        400
sports            400
Name: count, dtype: int64
```

Se definen las variables X e Y para el modelo

```
[8]: target_feature = 'category'
```

```
[9]: x_feature = 'content'
```

Se genera una copia de la información para no modificar la original para el proceso exploratorio de transformación de los datos:

```
[10]: X_train_trans = train_df.copy()
X_train_trans
```

```
[10]:                                                 headlines \
636   Rajinikanth fan mocks Vijay starrer The Greate...
161   Agastya Nanda says he probably didn't deserve ...
855   Malaikottai Valiban new poster out: Mohanlal i...
24    Hanu Man actor Teja Sajja on the responsibilit...
252   Arun Matheswaran: 'Captain Miller is my least ...
...
936   Women's World Cup: Deepti Sharma, Richa Ghosh ...
1378  Former state-level Punjab hockey player lifts ...
757   'I told Babar Azam and Saqlain Mushtaq to drop...
622   Watch: RB Leipzig's Benjamin Henrichs' handbal...
1629      Asia Cup set to be moved out of Pakistan
```

```
                                                 description \
636   A Rajinikanth fan shared a poster of Will Smit...
161   Agastya Nanda also revealed why he did not fee...
855   Lijo Jose Pellissery and Mohanlal have been ti...
24    Teja Sajja and Prasanth Varma's Hanu Man has p...
252   Arun Matheswaran calls Dhanush one of the shar...
...
936   Windies slump to 15th straight loss, eighth su...
```

```

1378 Kumar, who was part of Sports Authority of Ind...
757 Rizwan claimed in an interview to Cricket Paki...
622 The incident occurred late in the game after L...
1629 UAE could be the venue; BCCI ok with PCB hosti...

```

```

content \
636 Director Venkat Prabhu has never shied away fr...
161 Actor Agastya Nanda, who was recently seen in ...
855 If Kalki AD 2989 is the next big thing in the ...
24 Actor Teja Sajja's mythological film Hanu Man...
252 Speaking at the audio launch of Captain Miller...
...
936 With four needed to win, Richa Ghosh lined up ...
1378 He stands out like a sore thumb, as for some i...
757 Interesting developments across the border in ...
622 RB Leipzig's Benjamin Henrichs' handball incid...
1629 The Board of Control for Cricket in India (BCC...

```

	url	category
636	https://indianexpress.com/article/entertainmen...	entertainment
161	https://indianexpress.com/article/entertainmen...	entertainment
855	https://indianexpress.com/article/entertainmen...	entertainment
24	https://indianexpress.com/article/entertainmen...	entertainment
252	https://indianexpress.com/article/entertainmen...	entertainment
...
936	https://indianexpress.com/article/sports/crick...	sports
1378	https://indianexpress.com/article/sports/forme...	sports
757	https://indianexpress.com/article/sports/crick...	sports
622	https://indianexpress.com/article/sports/footb...	sports
1629	https://indianexpress.com/article/sports/crick...	sports

[8000 rows x 5 columns]

5.3 1.3. Exploración de los datos

Se va a generar un WordCloud para visualizar las palabras más frecuentes en las categorías.

Se inicia definiendo una función:

```

[11]: def show_wordcloud(palabras,stopwords=[]):
    comment_words = ''
    # iterate through the csv file
    for val in palabras:
        # typecaste each val to string
        val = str(val)

```

```

# split the value
tokens = val.split()

# Converts each token into lowercase
for i in range(len(tokens)):
    tokens[i] = tokens[i].lower()

comment_words += " ".join(tokens)+" "

wordcloud = WordCloud(width = 800, height = 800,
                      background_color ='white',
                      stopwords = stopwords,
                      min_font_size = 10).generate(comment_words)

# plot the WordCloud image
plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)

plt.show()

```

Se genera el llamada para cada una de las clases:

```
[12]: for i in train_df[target_feature].unique():
    print(f'----- Words for class: {i} -----')
    show_wordcloud(train_df.loc[train_df[target_feature]==i, x_feature])
```

----- Words for class: entertainment -----

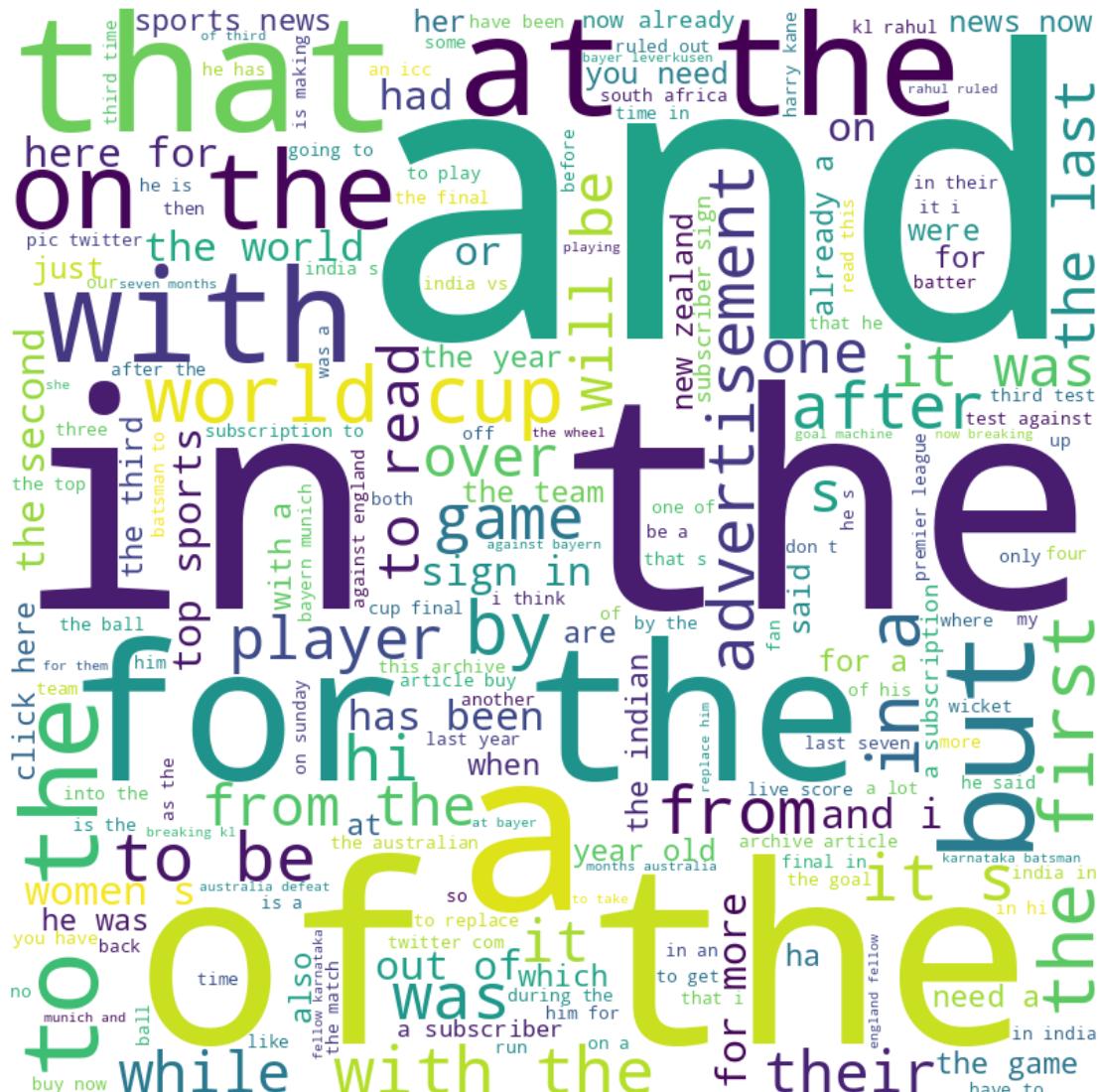
----- Words for class: education -----

----- Words for class: business -----

----- Words for class: technology -----



----- Words for class: sports -----

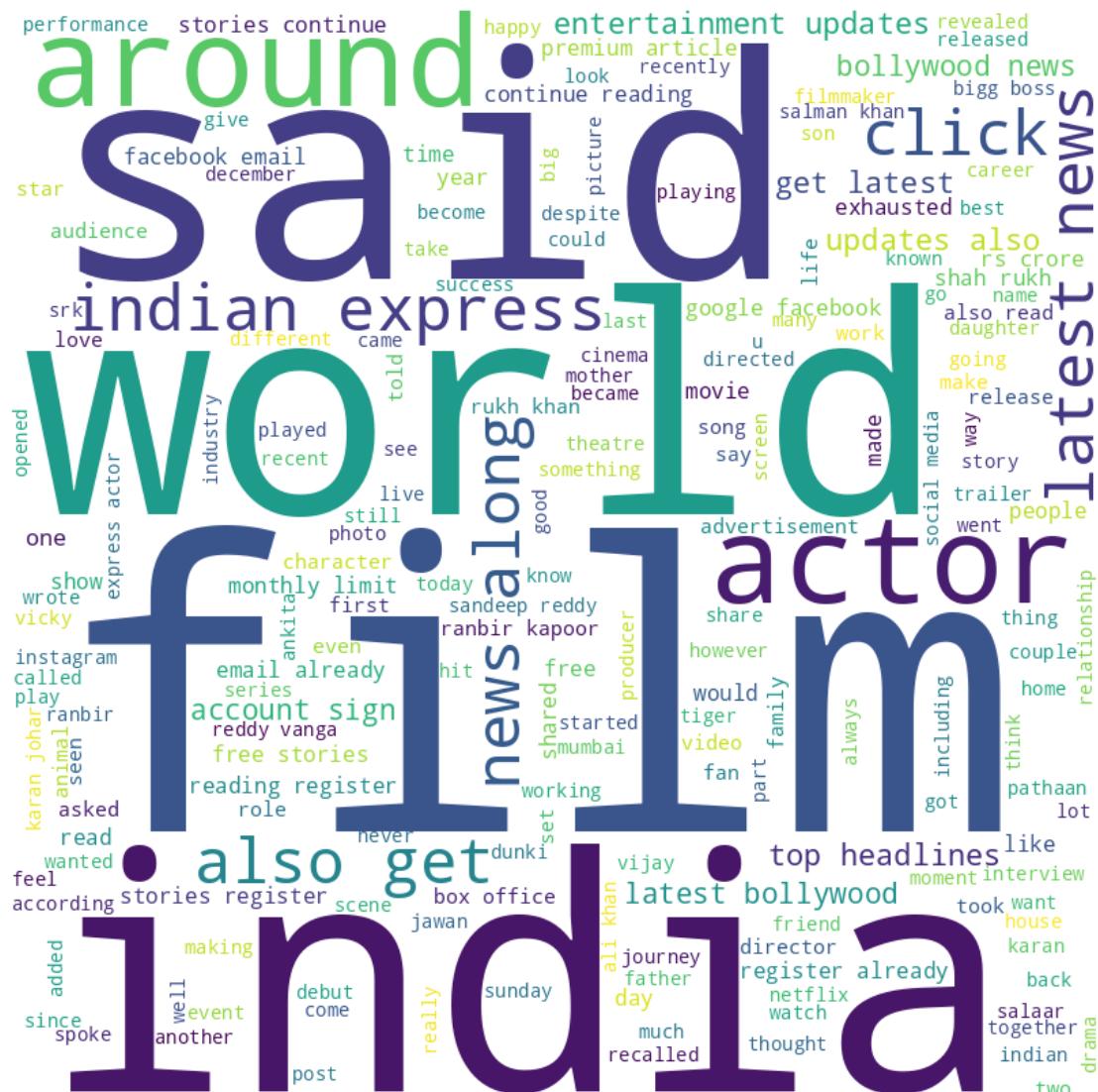


Como se puede ver, hay varias palabras que se repiten en las diferentes categorías, lo que puede generar ruido en el modelo, ya que no aportan mucha información, esas palabras se conocen como **stopwords**. Se va a proceder a eliminarlas y a realizar un nuevo WordCloud para visualizar las palabras más frecuentes en las categorías.

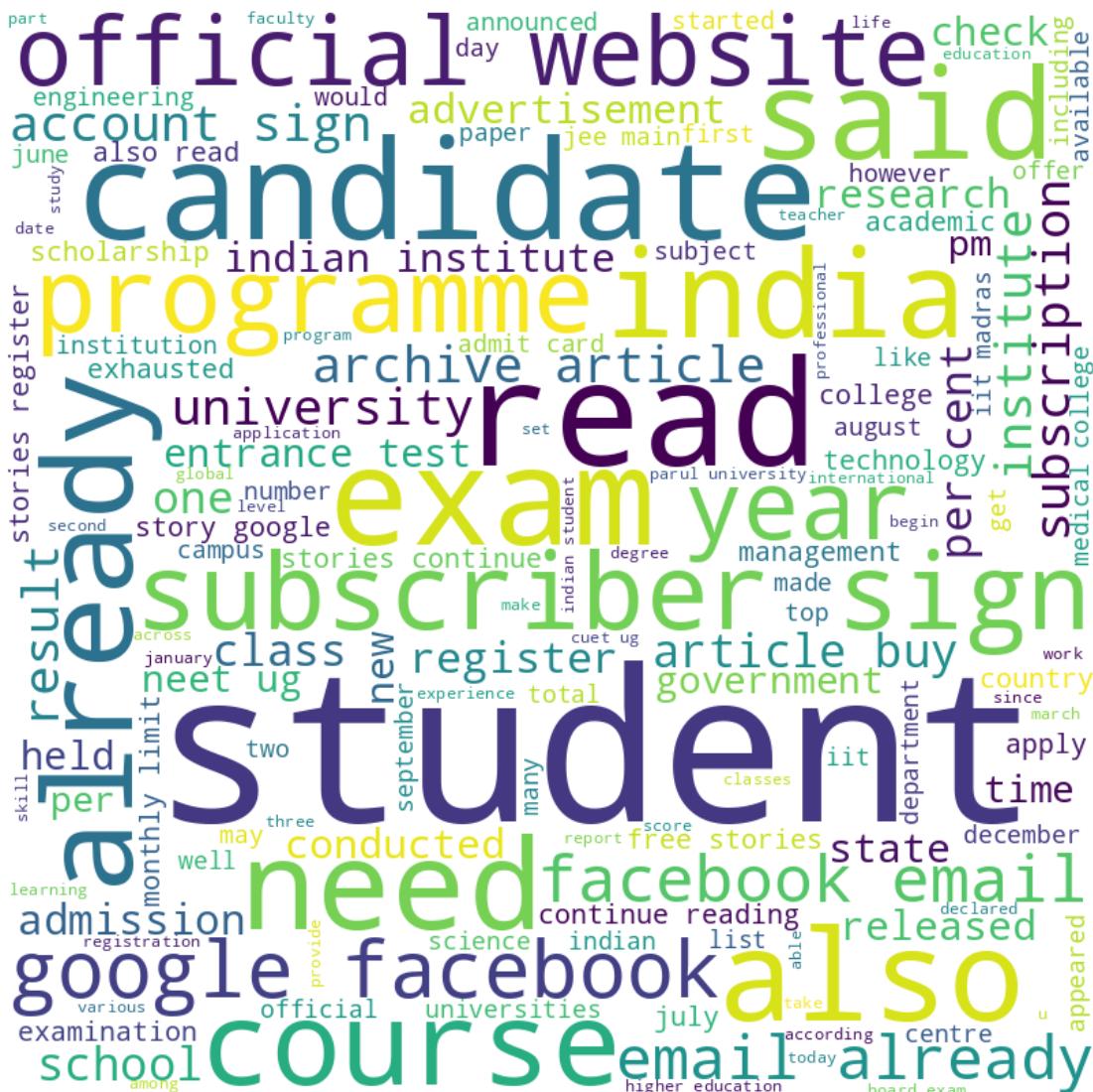
```
[13]: stop_words = stopwords.words('english')
```

```
[14]: for i in train_df[target_feature].unique():
    print(f'----- Words for class: {i} -----')
    show_wordcloud(train_df.loc[train_df[target_feature]==i, x_feature], □
                   stop_words)
```

----- Words for class: entertainment -----



----- Words for class: education -----



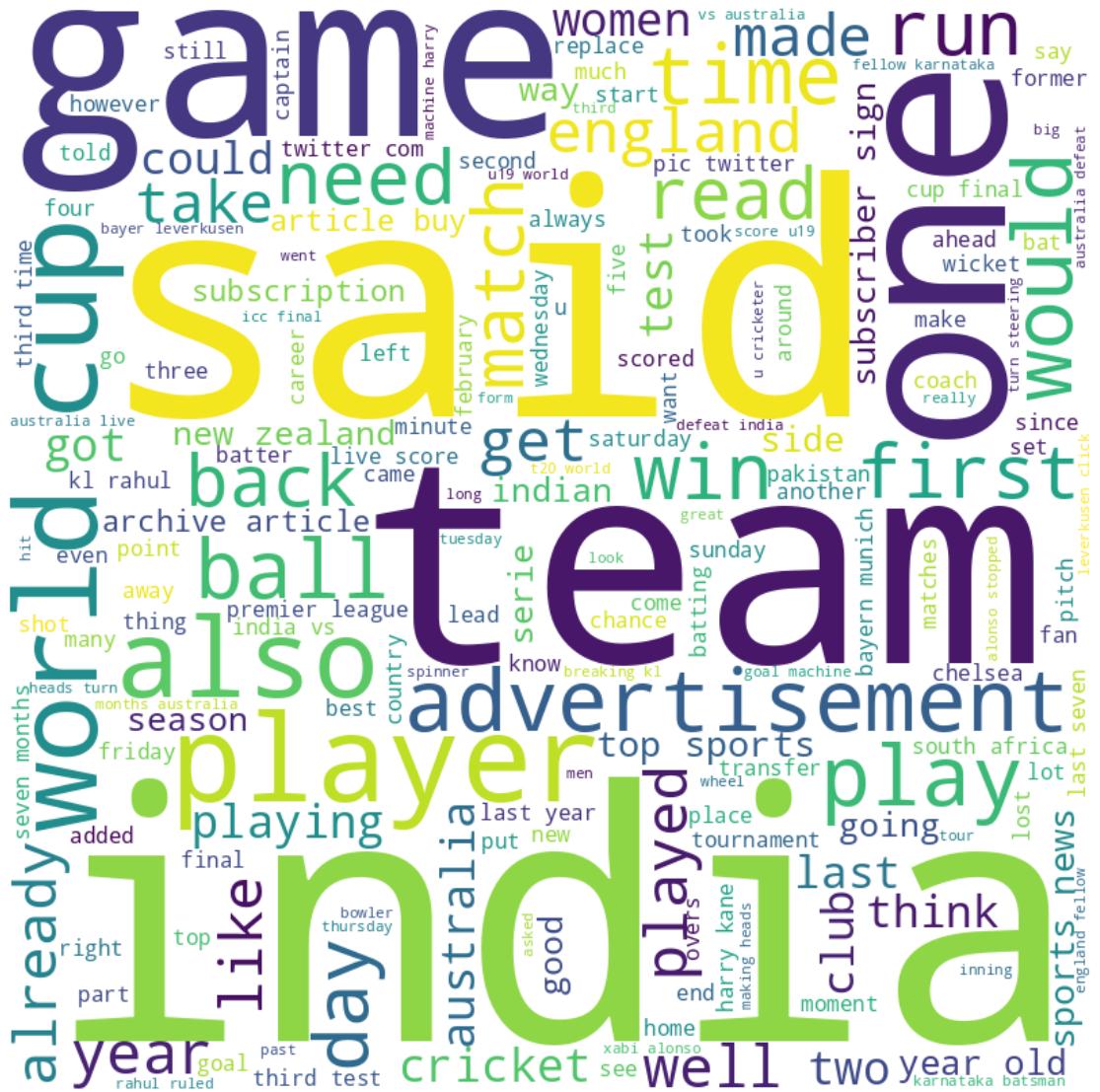
----- Words for class: business -----



----- Words for class: technology -----



----- Words for class: sports -----



Asimismo, es importante revisar que todas las palabras se encuentren en el mismo idioma dado que este proceso es sensible al idioma. Para ello, se toma la función `set_language` la cual utiliza la librería polyglot para reconocer en qué idioma se encuentra la mayoría de las filas. El resultado es que la mayoría se encuentra en: **inglés**.

```
[15]: import cld2

def set_language(val):
    # Remove invalid UTF-8 characters
    cleaned_text = val.encode('utf-8', 'ignore').decode('utf-8')

    # Detect language
    reliable, _, top_3_choices = cld2.detect(cleaned_text, bestEffort=False)
```

```

if reliable:
    return top_3_choices[0][1].lower()
else:
    return 'unknown'

train_df["language"] = train_df[x_feature].apply(set_language)
print(f"El lenguaje predominante es: {train_df['language'].unique()[0]}")

```

El lenguaje predominante es: en

Nota: Dado que el único que lenguaje que aparece es inglés no se eliminan registros.

5.4 1.4. Preparación de la información

Inicialmente, se separa tanto la variable objetivo como la variable independiente. Además, se convierte los valores targets en valores numéricos para que sean entendibles por el algoritmo.

```
[16]: label_encoder = LabelEncoder()
train_df[target_feature] = label_encoder.fit_transform(train_df[target_feature])
test_df[target_feature] = label_encoder.fit_transform(test_df[target_feature])

unique_labels = label_encoder.classes_
for num_value, original_label in enumerate(unique_labels):
    print(f'Valor numérico: {num_value}, Etiqueta original: {original_label}')

```

Valor numérico: 0, Etiqueta original: business
 Valor numérico: 1, Etiqueta original: education
 Valor numérico: 2, Etiqueta original: entertainment
 Valor numérico: 3, Etiqueta original: sports
 Valor numérico: 4, Etiqueta original: technology

Se realiza separación de *train*:

```
[17]: X_train, Y_train = train_df[x_feature], train_df[target_feature]
display(X_train)
Y_train
```

```

636     Director Venkat Prabhu has never shied away fr...
161     Actor Agastya Nanda, who was recently seen in ...
855     If Kalki AD 2989 is the next big thing in the ...
24      Actor Teja Sajja's mythological film Hanu Man,...
252     Speaking at the audio launch of Captain Miller...
...
936     With four needed to win, Richa Ghosh lined up ...
1378    He stands out like a sore thumb, as for some i...
757     Interesting developments across the border in ...
622     RB Leipzig's Benjamin Henrichs' handball incid...
1629    The Board of Control for Cricket in India (BCC...
Name: content, Length: 8000, dtype: object

```

```
[17]: 636    2  
161    2  
855    2  
24     2  
252    2  
..  
936    3  
1378   3  
757    3  
622    3  
1629   3  
Name: category, Length: 8000, dtype: int64
```

Se realiza separación de *test*:

```
[18]: X_test, Y_test = test_df[x_feature], test_df[target_feature]  
display(X_test)  
Y_test
```

```
321    The Internet Movie Database (IMDb) has unveile...  
1775   TV actors Divyanka Tripathi and Vivek Dahiya m...  
953    Director Jude Anthany Joseph took to social me...  
529    Deepika Padukone has always said that she and ...  
1878   Actor Mansoor Ali Khan's offensive and misogyn...  
...  
1006   Chess Grandmaster Alireza Firouzja, the Irania...  
1272   India vs New Zealand (IND vs NZ) 3rd T20I: Ind...  
1497   Indian wicket-keeper batter Dinesh Karthik bel...  
1756   Sania Mirza-Rohan Bopanna, Australian Open Mix...  
921    Should Australia play with two spinners or go ...  
Name: content, Length: 2000, dtype: object
```

```
[18]: 321    2  
1775   2  
953    2  
529    2  
1878   2  
..  
1006   3  
1272   3  
1497   3  
1756   3  
921    3  
Name: category, Length: 2000, dtype: int64
```

Se propone:

- Eliminación del Ruido.
- Tokenización.

- Normalización.

```
[19]: def remove_non_ascii(words):
    """Remove non-ASCII characters from list of tokenized words"""
    new_words = []
    for word in words:
        new_word = unicodedata.normalize('NFKD', word).encode('ascii', 'ignore').decode('utf-8', 'ignore')
        new_words.append(new_word)
    return new_words

def to_lowercase(words):
    """Convert all characters to lowercase from list of tokenized words"""
    new_words = []
    for word in words:
        new_word = word.lower()
        new_words.append(new_word)
    return new_words

def remove_punctuation(words):
    """Remove punctuation from list of tokenized words"""
    new_words = []
    for word in words:
        new_word = re.sub(r'[^w\s]', '', word)
        if new_word != '':
            new_words.append(new_word)
    return new_words

def replace_numbers(words):
    """Replace all interger occurrences in list of tokenized words with textual representation"""
    p = inflect.engine()
    new_words = []
    for word in words:
        if word.isdigit():
            new_word = p.number_to_words(word)
            new_words.append(new_word)
        else:
            new_words.append(word)
    return new_words

def remove_stopwords(words, stopwords=stopwords.words('english')):
    """Remove stop words from list of tokenized words"""
    new_words = []
    for word in words:
        if word not in stopwords:
            new_words.append(word)
```

```

    return new_words

def preprocessing(words):
    words = to_lowercase(words)
    words = replace_numbers(words)
    words = remove_punctuation(words)
    words = remove_non_ascii(words)
    words = remove_stopwords(words)
    return words

```

1.4.1 Tokenización Ahora se aplica la función a la columna X_feature y se aplica el preprocesamiento de los datos. Para ello se crea una función que realice las siguientes tareas:

- * Pasar a minúsculas.
- * Reemplazar los números por su correspondiente palabra.
- * Eliminar los signos de puntuación.
- * Eliminar caracteres especiales.
- * Eliminar stopwords.

[20]:

```
X_train_new = X_train.apply(word_tokenize)
X_train_new = X_train_new.apply(preprocessing) #Aplica la eliminación del ruido
X_train_new.head()
```

[20]:

```
636 [director, venkat, prabhu, never, shied, away, ...
161 [actor, agastya, nanda, recently, seen, zoya, ...
855 [kalki, ad, two thousand nine hundred and eigh...
24 [actor, teja, sajja, mythological, film, hanu, ...
252 [speaking, audio, launch, captain, miller, dha...
Name: content, dtype: object
```

[21]:

```
X_train_trans['trans'] = X_train_trans['content'].apply(nltk.
    word_tokenize,language="english").apply(preprocessing)
X_train_trans
```

[21]:

```
headlines \
636 Rajinikanth fan mocks Vijay starrer The Greate...
161 Agastya Nanda says he probably didn't deserve ...
855 Malaikottai Valiban new poster out: Mohanlal i...
24 Hanu Man actor Teja Sajja on the responsibilit...
252 Arun Matheswaran: 'Captain Miller is my least ...
...
936 Women's World Cup: Deepti Sharma, Richa Ghosh ...
1378 Former state-level Punjab hockey player lifts ...
757 'I told Babar Azam and Saqlain Mushtaq to drop...
622 Watch: RB Leipzig's Benjamin Henrichs' handbal...
1629 Asia Cup set to be moved out of Pakistan
```

```
description \
636 A Rajinikanth fan shared a poster of Will Smit...
161 Agastya Nanda also revealed why he did not fee...
855 Lijo Jose Pellissery and Mohanlal have been ti...
```

24 Teja Sajja and Prasanth Varma's Hanu Man has p...
252 Arun Matheswaran calls Dhanush one of the shar...
... ...
936 Windies slump to 15th straight loss, eighth su...
1378 Kumar, who was part of Sports Authority of Ind...
757 Rizwan claimed in an interview to Cricket Paki...
622 The incident occurred late in the game after L...
1629 UAE could be the venue; BCCI ok with PCB hosti...

content \

636 Director Venkat Prabhu has never shied away fr...
161 Actor Agastya Nanda, who was recently seen in ...
855 If Kalki AD 2989 is the next big thing in the ...
24 Actor Teja Sajja's mythological film Hanu Man,...
252 Speaking at the audio launch of Captain Miller...
... ...
936 With four needed to win, Richa Ghosh lined up ...
1378 He stands out like a sore thumb, as for some i...
757 Interesting developments across the border in ...
622 RB Leipzig's Benjamin Henrichs' handball incid...
1629 The Board of Control for Cricket in India (BCC...

url category \

636 <https://indianexpress.com/article/entertainmen...> entertainment
161 <https://indianexpress.com/article/entertainmen...> entertainment
855 <https://indianexpress.com/article/entertainmen...> entertainment
24 <https://indianexpress.com/article/entertainmen...> entertainment
252 <https://indianexpress.com/article/entertainmen...> entertainment
... ...
936 <https://indianexpress.com/article/sports/crick...> sports
1378 <https://indianexpress.com/article/sports/forme...> sports
757 <https://indianexpress.com/article/sports/crick...> sports
622 <https://indianexpress.com/article/sports/footb...> sports
1629 <https://indianexpress.com/article/sports/crick...> sports

trans

636 [director, venkat, prabhu, never, shied, away,...
161 [actor, agastya, nanda, recently, seen, zoya, ...
855 [kalki, ad, two thousand nine hundred and eigh...
24 [actor, teja, sajja, mythological, film, hanu,...
252 [speaking, audio, launch, captain, miller, dha...
... ...
936 [four, needed, win, richa, ghosh, lined, pull,...
1378 [stands, like, sore, thumb, inexplicable, reas...
757 [interesting, developments, across, border, pa...
622 [rb, leipzig, benjamin, henrichs, handball, in...
1629 [board, control, cricket, india, bcci, made, s...

```
[8000 rows x 6 columns]
```

1.4.2 Normalización Para la normalización de los datos se realiza una eliminación de prefijos y sufijos, además de realizar una lemmatización de los verbos.

```
[22]: def stem_words(words):
    """Stem words in list of tokenized words"""
    stemmer = SnowballStemmer('english')
    stems = []
    for word in words:
        stem = stemmer.stem(word)
        stems.append(stem)
    return stems

def lemmatize_verbs(words):
    """Lemmatize verbs in list of tokenized words"""
    lemmatizer = WordNetLemmatizer()
    lemmas = []
    for word in words:
        lemma = lemmatizer.lemmatize(word, pos='v')
        lemmas.append(lemma)
    return lemmas

def stem_and_lemmatize(words):
    words = stem_words(words)
    words = lemmatize_verbs(words)
    return words
```

```
[23]: X_train_new = X_train_new.apply(stem_and_lemmatize) #Aplica lematización y
        ↪Eliminación de Prefijos y Sufijos.
X_train_new.head()
```

```
[23]: 636 [director, venkat, prabhu, never, shi, away, a...
161 [actor, agastya, nanda, recent, see, zoya, akh...
855 [kalki, ad, two thousand nine hundred and eigh...
24 [actor, teja, sajja, mytholog, film, hanu, man...
252 [speak, audio, launch, captain, miller, dhanus...
Name: content, dtype: object
```

```
[24]: X_train_trans['trans'] = X_train_trans['trans'].apply(stem_words)
X_train_trans
```

```
[24]:                                                 headlines \
636 Rajinikanth fan mocks Vijay starrer The Greate...
161 Agastya Nanda says he probably didn't deserve ...
855 Malaikottai Valiban new poster out: Mohanlal i...
```

24 Hanu Man actor Teja Sajja on the responsibilit...
252 Arun Matheswaran: 'Captain Miller is my least ...
... ...
936 Women's World Cup: Deepti Sharma, Richa Ghosh ...
1378 Former state-level Punjab hockey player lifts ...
757 'I told Babar Azam and Saqlain Mushtaq to drop...
622 Watch: RB Leipzig's Benjamin Henrichs' handbal...
1629 Asia Cup set to be moved out of Pakistan

description \

636 A Rajinikanth fan shared a poster of Will Smit...
161 Agastya Nanda also revealed why he did not fee...
855 Lijo Jose Pellissery and Mohanlal have been ti...
24 Teja Sajja and Prasanth Varma's Hanu Man has p...
252 Arun Matheswaran calls Dhanush one of the shar...
... ...
936 Windies slump to 15th straight loss, eighth su...
1378 Kumar, who was part of Sports Authority of Ind...
757 Rizwan claimed in an interview to Cricket Paki...
622 The incident occurred late in the game after L...
1629 UAE could be the venue; BCCI ok with PCB hosti...

content \

636 Director Venkat Prabhu has never shied away fr...
161 Actor Agastya Nanda, who was recently seen in ...
855 If Kalki AD 2989 is the next big thing in the ...
24 Actor Teja Sajja's mythological film Hanu Man,...
252 Speaking at the audio launch of Captain Miller...
... ...
936 With four needed to win, Richa Ghosh lined up ...
1378 He stands out like a sore thumb, as for some i...
757 Interesting developments across the border in ...
622 RB Leipzig's Benjamin Henrichs' handball incid...
1629 The Board of Control for Cricket in India (BCC...

url category \

636 <https://indianexpress.com/article/entertainmen...> entertainment
161 <https://indianexpress.com/article/entertainmen...> entertainment
855 <https://indianexpress.com/article/entertainmen...> entertainment
24 <https://indianexpress.com/article/entertainmen...> entertainment
252 <https://indianexpress.com/article/entertainmen...> entertainment
... ...
936 <https://indianexpress.com/article/sports/crick...> sports
1378 <https://indianexpress.com/article/sports/forme...> sports
757 <https://indianexpress.com/article/sports/crick...> sports
622 <https://indianexpress.com/article/sports/footb...> sports
1629 <https://indianexpress.com/article/sports/crick...> sports

```

trans
636 [director, venkat, prabhu, never, shi, away, a...
161 [actor, agastya, nanda, recent, seen, zoya, ak...
855 [kalki, ad, two thousand nine hundred and eigh...
24 [actor, teja, sajja, mytholog, film, hanu, man...
252 [speak, audio, launch, captain, miller, dhanus...
...
...
936 [four, need, win, richa, ghosh, line, pull, sh...
1378 [stand, like, sore, thumb, inexplic, reason, p...
757 [interest, develop, across, border, pakistan, ...
622 [rb, leipzig, benjamin, henrich, handbal, inci...
1629 [board, control, cricket, india, bcci, made, s...

[8000 rows x 6 columns]

```

A continuación, se calculan algunas métricas:

```
[25]: X_train_trans['trans_count'] = X_train_trans['trans'].apply(lambda x: len(x))
X_train_trans
```

```

[25]:                                                 headlines \
636 Rajinikanth fan mocks Vijay starrer The Greate...
161 Agastya Nanda says he probably didn't deserve ...
855 Malaikottai Valiban new poster out: Mohanlal i...
24 Hanu Man actor Teja Sajja on the responsibilit...
252 Arun Matheswaran: 'Captain Miller is my least ...
...
...
936 Women's World Cup: Deepti Sharma, Richa Ghosh ...
1378 Former state-level Punjab hockey player lifts ...
757 'I told Babar Azam and Saqlain Mushtaq to drop...
622 Watch: RB Leipzig's Benjamin Henrichs' handbal...
1629 Asia Cup set to be moved out of Pakistan

                                                 description \
636 A Rajinikanth fan shared a poster of Will Smit...
161 Agastya Nanda also revealed why he did not fee...
855 Lijo Jose Pellissery and Mohanlal have been ti...
24 Teja Sajja and Prasanth Varma's Hanu Man has p...
252 Arun Matheswaran calls Dhanush one of the shar...
...
...
936 Windies slump to 15th straight loss, eighth su...
1378 Kumar, who was part of Sports Authority of Ind...
757 Rizwan claimed in an interview to Cricket Paki...
622 The incident occurred late in the game after L...
1629 UAE could be the venue; BCCI ok with PCB hosti...
```

	content \
636	Director Venkat Prabhu has never shied away fr...
161	Actor Agastya Nanda, who was recently seen in ...
855	If Kalki AD 2989 is the next big thing in the ...
24	Actor Teja Sajja's mythological film Hanu Man,...
252	Speaking at the audio launch of Captain Miller...
...	...
936	With four needed to win, Richa Ghosh lined up ...
1378	He stands out like a sore thumb, as for some i...
757	Interesting developments across the border in ...
622	RB Leipzig's Benjamin Henrichs' handball incid...
1629	The Board of Control for Cricket in India (BCC...

	url \ category
636	https://indianexpress.com/article/entertainmen... entertainment
161	https://indianexpress.com/article/entertainmen... entertainment
855	https://indianexpress.com/article/entertainmen... entertainment
24	https://indianexpress.com/article/entertainmen... entertainment
252	https://indianexpress.com/article/entertainmen... entertainment
...	...
936	https://indianexpress.com/article/sports/crick... sports
1378	https://indianexpress.com/article/sports/forme... sports
757	https://indianexpress.com/article/sports/crick... sports
622	https://indianexpress.com/article/sports/footb... sports
1629	https://indianexpress.com/article/sports/crick... sports

	trans trans_count
636	[director, venkat, prabhu, never, shi, away, a... 65
161	[actor, agastya, nanda, recent, seen, zoya, ak... 92
855	[kalki, ad, two thousand nine hundred and eigh... 109
24	[actor, teja, sajja, mytholog, film, hanu, man... 79
252	[speak, audio, launch, captain, miller, dhanus... 63
...	...
936	[four, need, win, richa, ghosh, line, pull, sh... 109
1378	[stand, like, sore, thumb, inexplic, reason, p... 54
757	[interest, develop, across, border, pakistan, .. 65
622	[rb, leipzig, benjamin, henrich, handbal, inci... 52
1629	[board, control, cricket, india, bcci, made, s... 43

[8000 rows x 7 columns]

```
[26]: print(f"El número promedio de tokens es: {X_train_trans['trans_count'].mean()}")
```

El número promedio de tokens es: 134.212625

Finalmente, se une para que en vez de ser una lista sea un string únicamente:

```
[27]: train_df['trans'] = X_train_new.apply(lambda x: ' '.join(map(str, x)))
train_df
```

```
[27]:                                                 headlines \
636 Rajinikanth fan mocks Vijay starrer The Greate...
161 Agastya Nanda says he probably didn't deserve ...
855 Malaikottai Valiban new poster out: Mohanlal i...
24 Hanu Man actor Teja Sajja on the responsibilit...
252 Arun Matheswaran: 'Captain Miller is my least ...
...
...
```

```
936 Women's World Cup: Deepti Sharma, Richa Ghosh ...
1378 Former state-level Punjab hockey player lifts ...
757 'I told Babar Azam and Saqlain Mushtaq to drop...
622 Watch: RB Leipzig's Benjamin Henrichs' handbal...
1629 Asia Cup set to be moved out of Pakistan
```

```
description \
636 A Rajinikanth fan shared a poster of Will Smit...
161 Agastya Nanda also revealed why he did not fee...
855 Lijo Jose Pellissery and Mohanlal have been ti...
24 Teja Sajja and Prasanth Varma's Hanu Man has p...
252 Arun Matheswaran calls Dhanush one of the shar...
...
...
```

```
936 Windies slump to 15th straight loss, eighth su...
1378 Kumar, who was part of Sports Authority of Ind...
757 Rizwan claimed in an interview to Cricket Paki...
622 The incident occurred late in the game after L...
1629 UAE could be the venue; BCCI ok with PCB hosti...
```

```
content \
636 Director Venkat Prabhu has never shied away fr...
161 Actor Agastya Nanda, who was recently seen in ...
855 If Kalki AD 2989 is the next big thing in the ...
24 Actor Teja Sajja's mythological film Hanu Man, ...
252 Speaking at the audio launch of Captain Miller...
...
...
```

```
936 With four needed to win, Richa Ghosh lined up ...
1378 He stands out like a sore thumb, as for some i...
757 Interesting developments across the border in ...
622 RB Leipzig's Benjamin Henrichs' handball incid...
1629 The Board of Control for Cricket in India (BCC...
```

		url	category	language	\
636	https://indianexpress.com/article/entertainmen...	2	en		
161	https://indianexpress.com/article/entertainmen...	2	en		
855	https://indianexpress.com/article/entertainmen...	2	en		
24	https://indianexpress.com/article/entertainmen...	2	en		

			2	en
252	https://indianexpress.com/article/entertainmen...
...				
936	https://indianexpress.com/article/sports/crick...	...	3	en
1378	https://indianexpress.com/article/sports/forme...	...	3	en
757	https://indianexpress.com/article/sports/crick...	...	3	en
622	https://indianexpress.com/article/sports/footb...	...	3	en
1629	https://indianexpress.com/article/sports/crick...	...	3	en
		trans		
636	director venkat prabhu never shi away accept f...	...		
161	actor agastya nanda recent see zoya akhtar arc...	...		
855	kalki ad two thousand nine hundred and eighty...	...		
24	actor teja sajja mytholog film hanu man garner...	...		
252	speak audio launch captain miller dhanush say		
...		...		
936	four need win richa ghosh line pull shamilia c...	...		
1378	stand like sore thumb inexplic reason promis p...	...		
757	interest develop across border pakistan cricke...	...		
622	rb leipzig benjamin henrich handbal incid crea...	...		
1629	board control cricket india bcci make stand cl...	...		

[8000 rows x 7 columns]

5.4.1 1.5 Aplicación del Mismo Procesamiento a los Datos de Test

```
[28]: X_test_trans = test_df.copy()
X_test_trans
```

```
[28]: headlines \
321 IMDb unveils list of most anticipated Indian m...
1775 Vivek Dahiya was apprehensive about marrying D...
953 2018 director Jude Anthany Joseph apologises a...
529 Deepika Padukone says she's looking forward to...
1878 Mansoor Ali Khan's remarks about Trisha are pa...
...
1006 Chess Grandmaster Alireza Firouzja forays into...
1272 India vs New Zealand (IND vs NZ) 3rd T20I Live...
1497 Border-Gavaskar Trophy: Dinesh Karthik picks V...
1756 Sania Mirza, Rohan Bopanna's Australian Open M...
921 IND vs AUS: Australia should 'play three seame...
```

```
description \
321 According to the website, the list was compile...
1775 When Vivek Dahiya was suggested the idea of ma...
953 2018 director Jude Anthany Joseph has reacted ...
529 In a new interview, Deepika Padukone shared th...
1878 From Kamal Haasan kissing minor Rekha on-scre...
```

...
 1006 Chess GM Alireza Firouzja says he has been in ...
 1272 IND vs NZ 3rd T20I Live: When & Where To Watch...
 1497 Kohli has scored 1682 runs in 20 Test matches ...
 1756 Sania Mirza and Rohan Bopanna will be competin...
 921 And which spinner that would be, considering T...

content \

321 The Internet Movie Database (IMDb) has unveile...
 1775 TV actors Divyanka Tripathi and Vivek Dahiya m...
 953 Director Jude Anthany Joseph took to social me...
 529 Deepika Padukone has always said that she and ...
 1878 Actor Mansoor Ali Khan's offensive and misogyn...
 ...
 1006 Chess Grandmaster Alireza Firouzja, the Irania...
 1272 India vs New Zealand (IND vs NZ) 3rd T20I: Ind...
 1497 Indian wicket-keeper batter Dinesh Karthik bel...
 1756 Sania Mirza-Rohan Bopanna, Australian Open Mix...
 921 Should Australia play with two spinners or go ...

	url	category
321	https://indianexpress.com/article/entertainmen...	2
1775	https://indianexpress.com/article/entertainmen...	2
953	https://indianexpress.com/article/entertainmen...	2
529	https://indianexpress.com/article/entertainmen...	2
1878	https://indianexpress.com/article/entertainmen...	2
...
1006	https://indianexpress.com/article/sports/chess...	3
1272	https://indianexpress.com/article/sports/crick...	3
1497	https://indianexpress.com/article/sports/crick...	3
1756	https://indianexpress.com/article/sports/tenni...	3
921	https://indianexpress.com/article/sports/crick...	3

[2000 rows x 5 columns]

```
[29]: X_test_new = X_train.apply(word_tokenize)
X_test_new = X_train_new.apply(preprocessing) #Aplica la eliminación del ruido
X_test_new.head()
```

```
[29]: 636 [director, venkat, prabhu, never, shi, away, a...
161 [actor, agastya, nanda, recent, see, zoya, akh...
855 [kalki, ad, two thousand nine hundred and eigh...
24 [actor, teja, sajja, mytholog, film, hanu, man...
252 [speak, audio, launch, captain, miller, dhanus...
Name: content, dtype: object
```

```
[30]: X_test_trans['trans'] = X_test_trans['content'].apply(nltk.  
    ↪word_tokenize,language="english").apply(preprocessing)  
X_test_trans
```

```
[30]:                                                 headlines \  
321    IMDb unveils list of most anticipated Indian m...  
1775    Vivek Dahiya was apprehensive about marrying D...  
953    2018 director Jude Anthany Joseph apologises a...  
529    Deepika Padukone says she's looking forward to...  
1878    Mansoor Ali Khan's remarks about Trisha are pa...  
...          ...  
1006    Chess Grandmaster Alireza Firouzja forays into...  
1272    India vs New Zealand (IND vs NZ) 3rd T20I Live...  
1497    Border-Gavaskar Trophy: Dinesh Karthik picks V...  
1756    Sania Mirza, Rohan Bopanna's Australian Open M...  
921    IND vs AUS: Australia should 'play three seame...  
  
                                                 description \  
321    According to the website, the list was compile...  
1775    When Vivek Dahiya was suggested the idea of ma...  
953    2018 director Jude Anthany Joseph has reacted ...  
529    In a new interview, Deepika Padukone shared th...  
1878    From Kamal Haasan kissing minor Rekha on-scree...  
...          ...  
1006    Chess GM Alireza Firouzja says he has been in ...  
1272    IND vs NZ 3rd T20I Live: When & Where To Watch...  
1497    Kohli has scored 1682 runs in 20 Test matches ...  
1756    Sania Mirza and Rohan Bopanna will be competin...  
921    And which spinner that would be, considering T...  
  
                                                 content \  
321    The Internet Movie Database (IMDb) has unveile...  
1775    TV actors Divyanka Tripathi and Vivek Dahiya m...  
953    Director Jude Anthany Joseph took to social me...  
529    Deepika Padukone has always said that she and ...  
1878    Actor Mansoor Ali Khan's offensive and misogyn...  
...          ...  
1006    Chess Grandmaster Alireza Firouzja, the Irania...  
1272    India vs New Zealand (IND vs NZ) 3rd T20I: Ind...  
1497    Indian wicket-keeper batter Dinesh Karthik bel...  
1756    Sania Mirza-Rohan Bopanna, Australian Open Mix...  
921    Should Australia play with two spinners or go ...  
  
                                                 url   category \  
321    https://indianexpress.com/article/entertainmen...      2  
1775    https://indianexpress.com/article/entertainmen...      2  
953    https://indianexpress.com/article/entertainmen...      2
```

529	https://indianexpress.com/article/entertainmen...	2
1878	https://indianexpress.com/article/entertainmen...	2
...
1006	https://indianexpress.com/article/sports/chess..	3
1272	https://indianexpress.com/article/sports/crick..	3
1497	https://indianexpress.com/article/sports/crick..	3
1756	https://indianexpress.com/article/sports/tenni..	3
921	https://indianexpress.com/article/sports/crick..	3

	trans
321	[internet, movie, database, imdb, unveiled, li..
1775	[tv, actors, divyanka, tripathi, vivek, dahiya..
953	[director, jude, anthany, joseph, took, social..
529	[deepika, padukone, always, said, husband, act..
1878	[actor, mansoor, ali, khan, offensive, misogyn..
...	...
1006	[chess, grandmaster, alireza, firouzja, irania..
1272	[india, vs, new, zealand, ind, vs, nz, 3rd, t2..
1497	[indian, wicketkeeper, batter, dinesh, karthik..
1756	[sania, mirzarchan, bopanna, australian, open,..
921	[australia, play, two, spinners, go, three, se..

[2000 rows x 6 columns]

```
[31]: X_test_new = X_test_new.apply(stem_and_lemmatize) #Aplica lematización y
      ↵Eliminación de Prefijos y Sufijos.
X_test_new.head()
```

```
[31]: 636 [director, venkat, prabhu, never, shi, away, a...
  161 [actor, agastya, nanda, recent, see, zoya, akh...
  855 [kalki, ad, two thousand nine hundred and eigh...
   24 [actor, teja, sajja, mytholog, film, hanu, man...
  252 [speak, audio, launch, captain, miller, dhanus...
Name: content, dtype: object
```

```
[32]: X_test_trans['trans'] = X_test_trans['trans'].apply(stem_words)
X_test_trans
```

```
[32]:                                                 headlines \
321 IMDb unveils list of most anticipated Indian m...
1775 Vivek Dahiya was apprehensive about marrying D...
953 2018 director Jude Anthany Joseph apologises a...
529 Deepika Padukone says she's looking forward to...
1878 Mansoor Ali Khan's remarks about Trisha are pa...
...
1006 Chess Grandmaster Alireza Firouzja forays into...
1272 India vs New Zealand (IND vs NZ) 3rd T20I Live...
```

1497 Border-Gavaskar Trophy: Dinesh Karthik picks V...
1756 Sania Mirza, Rohan Bopanna's Australian Open M...
921 IND vs AUS: Australia should 'play three seame...

description \

321 According to the website, the list was compile...
1775 When Vivek Dahiya was suggested the idea of ma...
953 2018 director Jude Anthany Joseph has reacted ...
529 In a new interview, Deepika Padukone shared th...
1878 From Kamal Haasan kissing minor Rekha on-scree...
...
1006 Chess GM Alireza Firouzja says he has been in ...
1272 IND vs NZ 3rd T20I Live: When & Where To Watch...
1497 Kohli has scored 1682 runs in 20 Test matches ...
1756 Sania Mirza and Rohan Bopanna will be competin...
921 And which spinner that would be, considering T...

content \

321 The Internet Movie Database (IMDb) has unveile...
1775 TV actors Divyanka Tripathi and Vivek Dahiya m...
953 Director Jude Anthany Joseph took to social me...
529 Deepika Padukone has always said that she and ...
1878 Actor Mansoor Ali Khan's offensive and misogyn...
...
1006 Chess Grandmaster Alireza Firouzja, the Irania...
1272 India vs New Zealand (IND vs NZ) 3rd T20I: Ind...
1497 Indian wicket-keeper batter Dinesh Karthik bel...
1756 Sania Mirza-Rohan Bopanna, Australian Open Mix...
921 Should Australia play with two spinners or go ...

url category \

321 <https://indianexpress.com/article/entertainmen...> 2
1775 <https://indianexpress.com/article/entertainmen...> 2
953 <https://indianexpress.com/article/entertainmen...> 2
529 <https://indianexpress.com/article/entertainmen...> 2
1878 <https://indianexpress.com/article/entertainmen...> 2
...
1006 <https://indianexpress.com/article/sports/chess...> 3
1272 <https://indianexpress.com/article/sports/crick...> 3
1497 <https://indianexpress.com/article/sports/crick...> 3
1756 <https://indianexpress.com/article/sports/tenni...> 3
921 <https://indianexpress.com/article/sports/crick...> 3

trans

321 [internet, movi, databas, imdb, unveil, list, ...
1775 [tv, actor, divyanka, tripathi, vivek, dahiya,...
953 [director, jude, anthani, joseph, took, social...

```

529 [deepika, padukon, alway, said, husband, actor...
1878 [actor, mansoor, ali, khan, offens, misogynist...
...
1006 [chess, grandmast, alireza, firouzja, iranian...
1272 [india, vs, new, zealand, ind, vs, nz, 3rd, t2...
1497 [indian, wicketkeep, batter, dinesh, karthik, ...
1756 [sania, mirzarohan, bopanna, australian, open...
921 [australia, play, two, spinner, go, three, sea...

```

[2000 rows x 6 columns]

```

[33]: X_test_new.reset_index(drop=True, inplace=True)
test_df['trans'] = X_test_new.apply(lambda x: ' '.join(map(str, x)))
test_df

```

```

[33]:
headlines \
321 IMDb unveils list of most anticipated Indian m...
1775 Vivek Dahiya was apprehensive about marrying D...
953 2018 director Jude Anthany Joseph apologises a...
529 Deepika Padukone says she's looking forward to...
1878 Mansoor Ali Khan's remarks about Trisha are pa...
...
1006 Chess Grandmaster Alireza Firouzja forays into...
1272 India vs New Zealand (IND vs NZ) 3rd T20I Live...
1497 Border-Gavaskar Trophy: Dinesh Karthik picks V...
1756 Sania Mirza, Rohan Bopanna's Australian Open M...
921 IND vs AUS: Australia should 'play three seame...

```

```

description \
321 According to the website, the list was compile...
1775 When Vivek Dahiya was suggested the idea of ma...
953 2018 director Jude Anthany Joseph has reacted ...
529 In a new interview, Deepika Padukone shared th...
1878 From Kamal Haasan kissing minor Rekha on-scre...
...
1006 Chess GM Alireza Firouzja says he has been in ...
1272 IND vs NZ 3rd T20I Live: When & Where To Watch...
1497 Kohli has scored 1682 runs in 20 Test matches ...
1756 Sania Mirza and Rohan Bopanna will be competin...
921 And which spinner that would be, considering T...

```

```

content \
321 The Internet Movie Database (IMDb) has unveile...
1775 TV actors Divyanka Tripathi and Vivek Dahiya m...
953 Director Jude Anthany Joseph took to social me...
529 Deepika Padukone has always said that she and ...
1878 Actor Mansoor Ali Khan's offensive and misogyn...

```

...

1006	Chess Grandmaster Alireza Firouzja, the Irania...	...
1272	India vs New Zealand (IND vs NZ) 3rd T20I: Ind...	
1497	Indian wicket-keeper batter Dinesh Karthik bel...	
1756	Sania Mirza-Rohan Bopanna, Australian Open Mix...	
921	Should Australia play with two spinners or go ...	

	url	category	\
321	https://indianexpress.com/article/entertainmen...	2	
1775	https://indianexpress.com/article/entertainmen...	2	
953	https://indianexpress.com/article/entertainmen...	2	
529	https://indianexpress.com/article/entertainmen...	2	
1878	https://indianexpress.com/article/entertainmen...	2	

...
1006	https://indianexpress.com/article/sports/chess...	3	
1272	https://indianexpress.com/article/sports/crick...	3	
1497	https://indianexpress.com/article/sports/crick...	3	
1756	https://indianexpress.com/article/sports/tenni...	3	
921	https://indianexpress.com/article/sports/crick...	3	

	trans
321	sriram raghavan merri christma star katrina ka...
1775	nation test agenc nta relea ignou jat two thou...
953	much say ar rahman style work music compo best...
529	bollywood playback singer shaan thursday day d...
1878	govern nod two australian univ univ wollongong...

...	...
1006	director rohit shetti address alleg singham fi...
1272	sister janhvi kapoor khushi kapoor appear gues...
1497	reacher back well world soon three pal whose b...
1756	jharkhand board 12th art commerc result two th...
921	indian stream space reach viewer sinc pandem a...

[2000 rows x 6 columns]

5.4.2 1.6. Codificación de Texto

1.6.1 Codificación del Texto CountVectorizer

```
[34]: def tokenizer(text):
    return word_tokenize(text, language="english")

dummy = CountVectorizer(tokenizer=tokenizer, stop_words=stopwords.
    ↪words('english'), lowercase=True)
X_train_Bow = dummy.fit_transform(train_df['trans'])
X_test_Bow = dummy.transform(test_df['trans'])
```

/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine

```

Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/sklearn/feature_extraction/text.py:525: UserWarning: The parameter
'token_pattern' will not be used since 'tokenizer' is not None'
    warnings.warn(
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/sklearn/feature_extraction/text.py:408: UserWarning: Your stop_words
may be inconsistent with your preprocessing. Tokenizing the stop words generated
tokens ["'d", "'ll", "'re", "'s", "'ve", 'could', 'might', 'must', "n't",
'need', 'sha', 'wo', 'would'] not in stop_words.
    warnings.warn(

```

1.6.2 Codificación del Texto Tfifdf

[35]:

```

vectorizer = TfidfVectorizer()
X_train_TFID = vectorizer.fit_transform(train_df['trans'])
X_test_TFID = vectorizer.transform(test_df['trans'])

```

[38]:

```
X_train_TFID.shape
```

[38]:

```
(8000, 43713)
```

[39]:

```
X_test_TFID.shape
```

[39]:

```
(2000, 43713)
```

[36]:

```
scipy.sparse.issparse(X_train_TFID)
```

[36]:

```
True
```

Las categorías del proceso de vectorización son las siguientes:

[40]:

```

terms = vectorizer.get_feature_names_out()
print(f"El número de columnas es: {len(terms)}")
terms

```

El número de columnas es: 43713

[40]:

```
array(['00', '000', '001', ..., 'zverev', 'zwischenahn', 'zyada'],
      dtype=object)
```

[41]:

```
tfidf_df = pd.DataFrame(X_train_TFID.toarray(), columns=terms)
tfidf_df
```

[41]:

	00	000	001	002	003	004	005	006	007	008	...	zuckerberg	\
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	

```

4      0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...
...   ...  ...
7995  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...
7996  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...
7997  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...
7998  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...
7999  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...

      zuckerberg1  zulfon  zulili  zulkifli  zurich  zve10  zverev \
0          0.0     0.0     0.0     0.0     0.0     0.0     0.0
1          0.0     0.0     0.0     0.0     0.0     0.0     0.0
2          0.0     0.0     0.0     0.0     0.0     0.0     0.0
3          0.0     0.0     0.0     0.0     0.0     0.0     0.0
4          0.0     0.0     0.0     0.0     0.0     0.0     0.0
...
...   ...
7995  0.0     0.0     0.0     0.0     0.0     0.0     0.0
7996  0.0     0.0     0.0     0.0     0.0     0.0     0.0
7997  0.0     0.0     0.0     0.0     0.0     0.0     0.0
7998  0.0     0.0     0.0     0.0     0.0     0.0     0.0
7999  0.0     0.0     0.0     0.0     0.0     0.0     0.0

      zwischenahn  zyada
0          0.0     0.0
1          0.0     0.0
2          0.0     0.0
3          0.0     0.0
4          0.0     0.0
...
...   ...
7995  0.0     0.0
7996  0.0     0.0
7997  0.0     0.0
7998  0.0     0.0
7999  0.0     0.0

```

[8000 rows x 43713 columns]

5.5 Punto 1 y 2

Se define una función para visualizar los componentes principales:

```
[42]: #La función grafica el último de los componentes identificados con sus
       ↪respectivas clases
def draw_components(labels, X, Y, n_components):
    # Inicializar LSA (TruncatedSVD), similar a PCA pero para matrices dispersas
    pca = TruncatedSVD(n_components=n_components)

    if n_components < 2:
```

```

raise("El número de componentes no puede ser menor a 2")

# Ajustar y transformar los datos TF-IDF
X_pca = pca.fit_transform(X)
print(X_pca.shape)

print(sum(pca.explained_variance_ratio_))

#Paleta de colores
colors = plt.cm.viridis(np.linspace(0, 1, len(labels)))
label_color_dict = dict(zip(labels, colors))

# Asignar un color a cada etiqueta
label_colors = [label_color_dict[label_encoder.inverse_transform([label])[0]]]
for label in Y

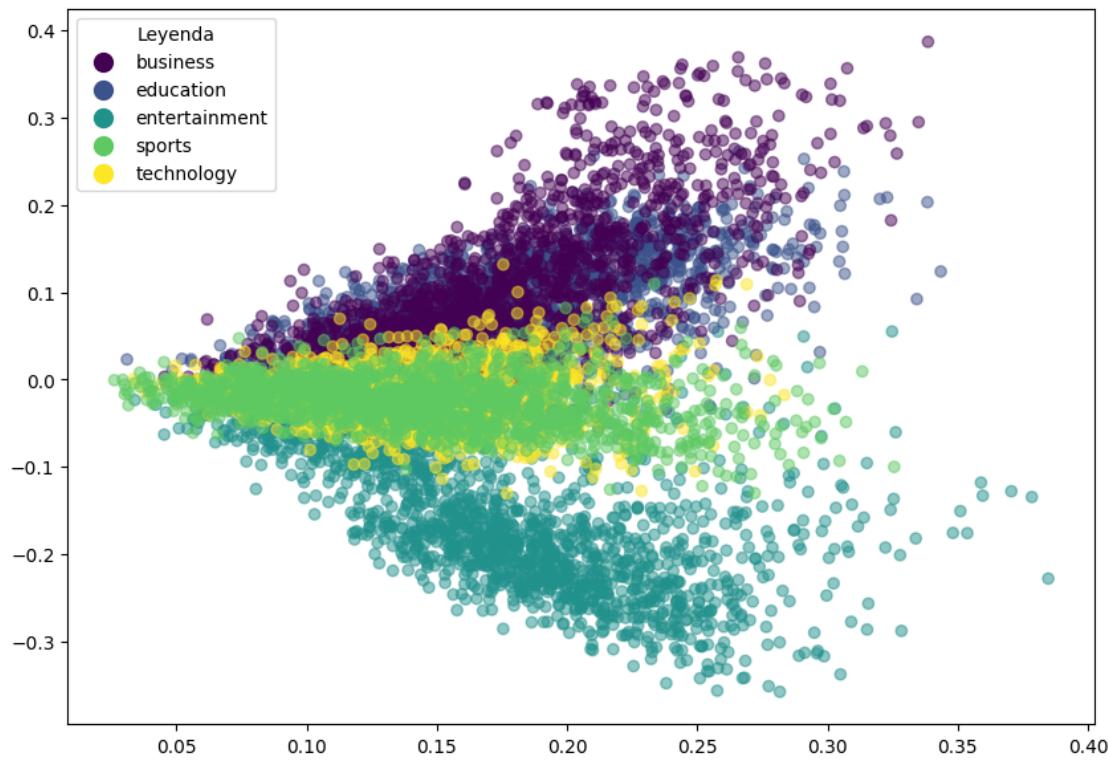
# Gráfico
plt.figure(figsize=(10, 7))
scatter = plt.scatter(X_pca[:, 0], X_pca[:, n_components-1], c=label_colors, alpha=0.5)

#Leyenda
handles = [plt.Line2D([0], [0], marker='o', color=color, linewidth=0, markersize=10) for label, color in label_color_dict.items()]
plt.legend(handles, labels, title='Leyenda')
plt.show()

```

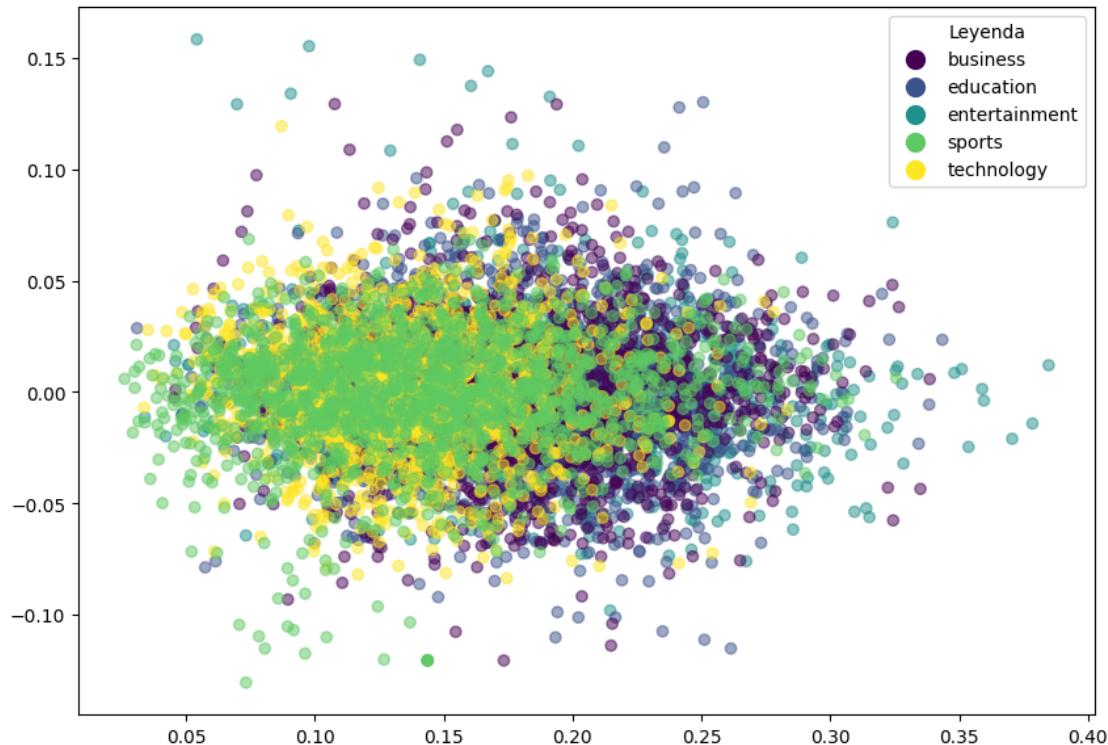
[46]: #Veamos gráficamente las componentes 1 y 2.
draw_components(unique_labels, tfidf_df, Y_train, 2)

(8000, 2)
0.016128134006590588



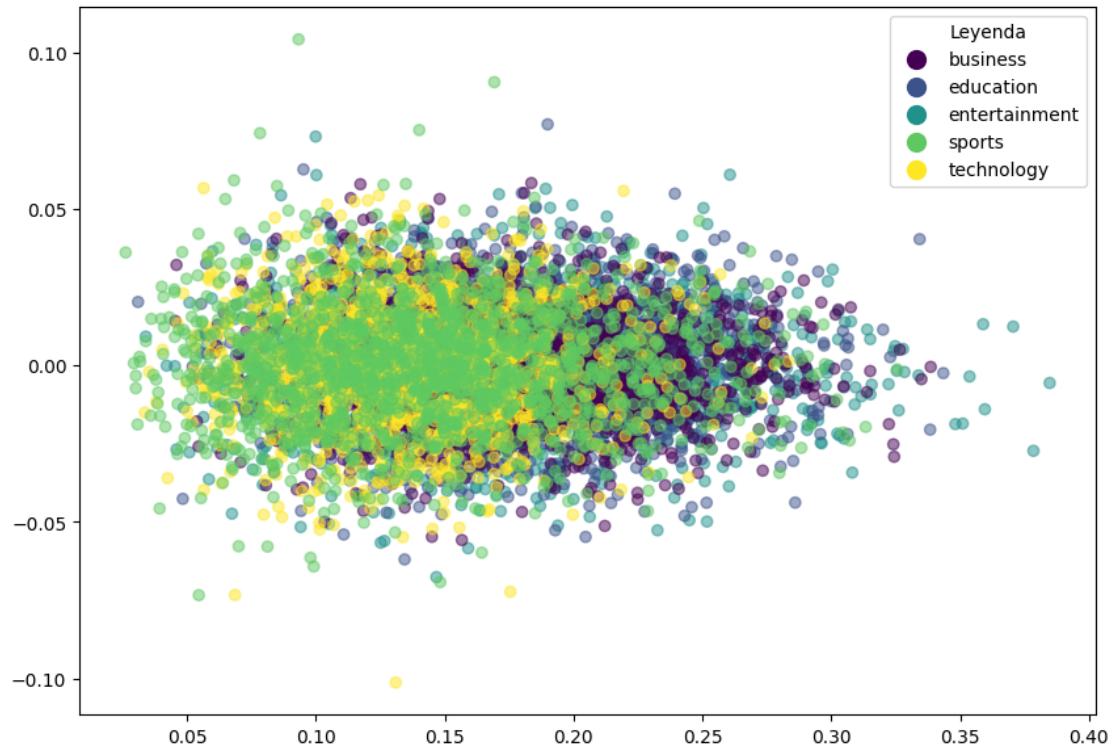
```
[47]: draw_components(unique_labels, tfidf_df, Y_train, 100)
```

```
(8000, 100)  
0.20883447897424398
```



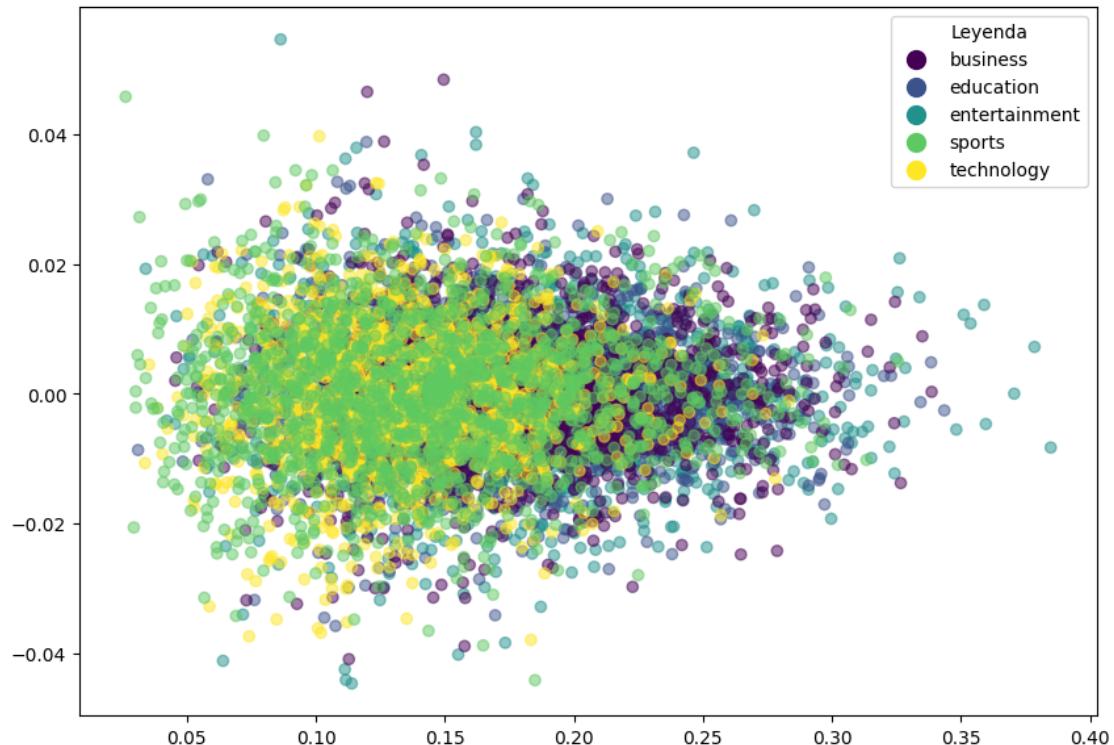
```
[48]: draw_components(unique_labels, tfidf_df, Y_train, 500)
```

```
(8000, 500)  
0.4265524403272992
```



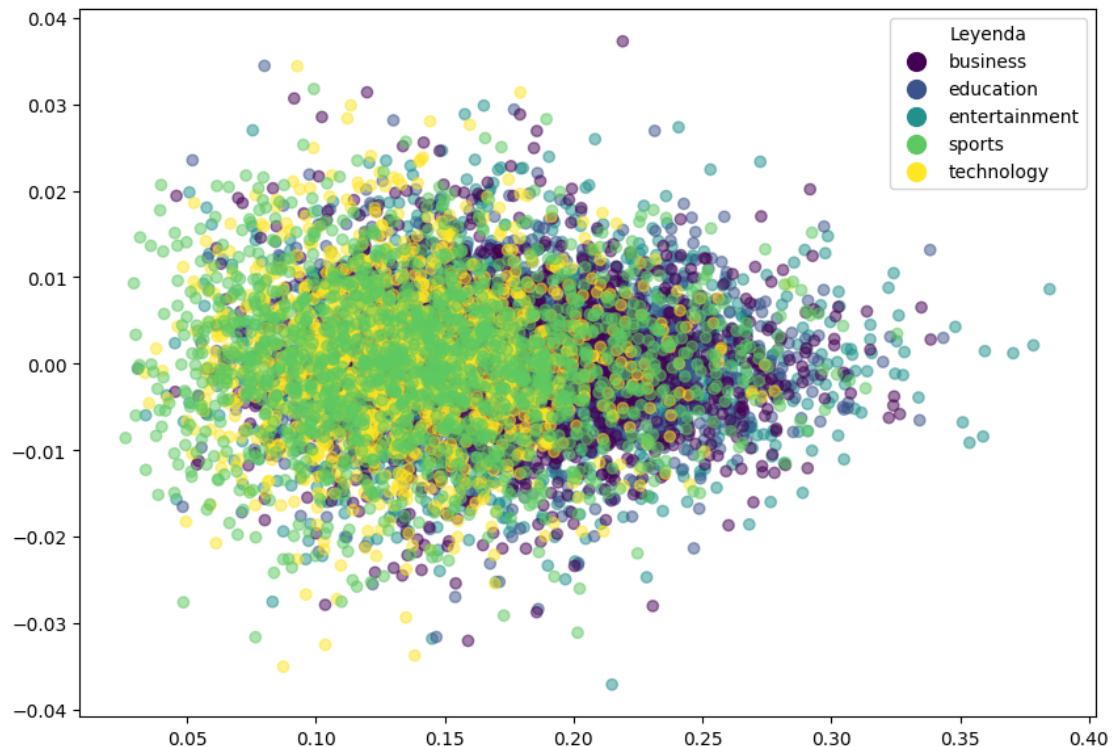
```
[49]: draw_components(unique_labels, tfidf_df, Y_train, 2000)
```

```
(8000, 2000)
0.7397817924101162
```



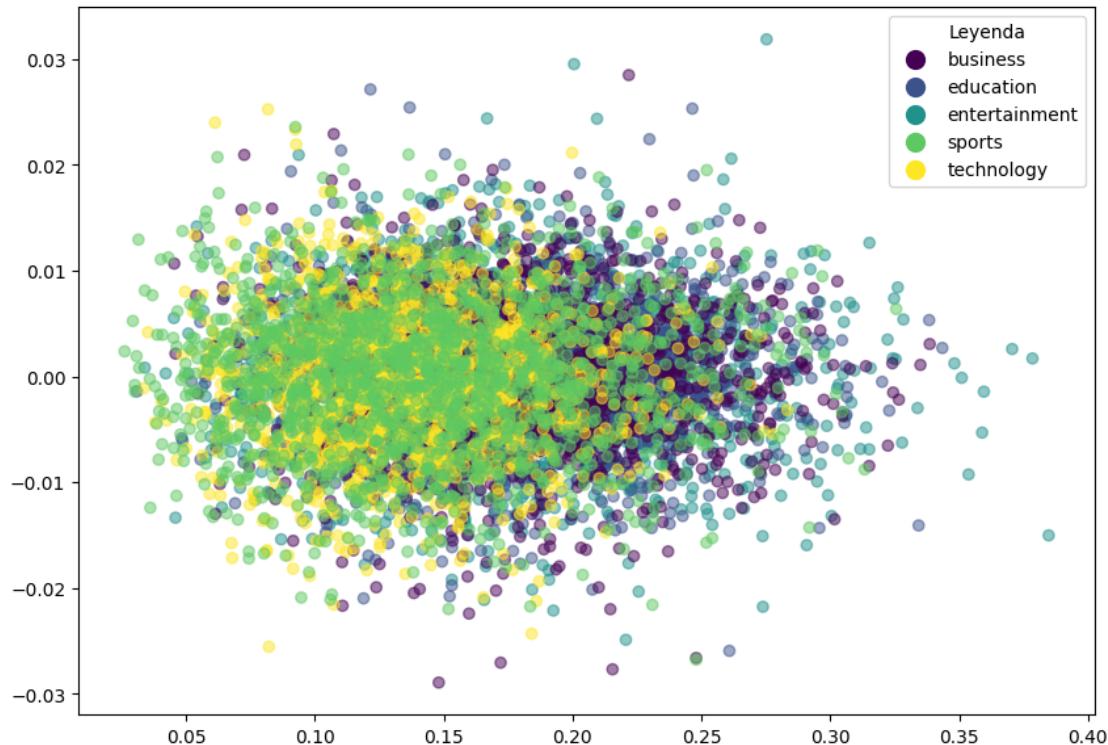
```
[50]: draw_components(unique_labels, tfidf_df, Y_train, 3000)
```

```
(8000, 3000)
0.8384314324158069
```



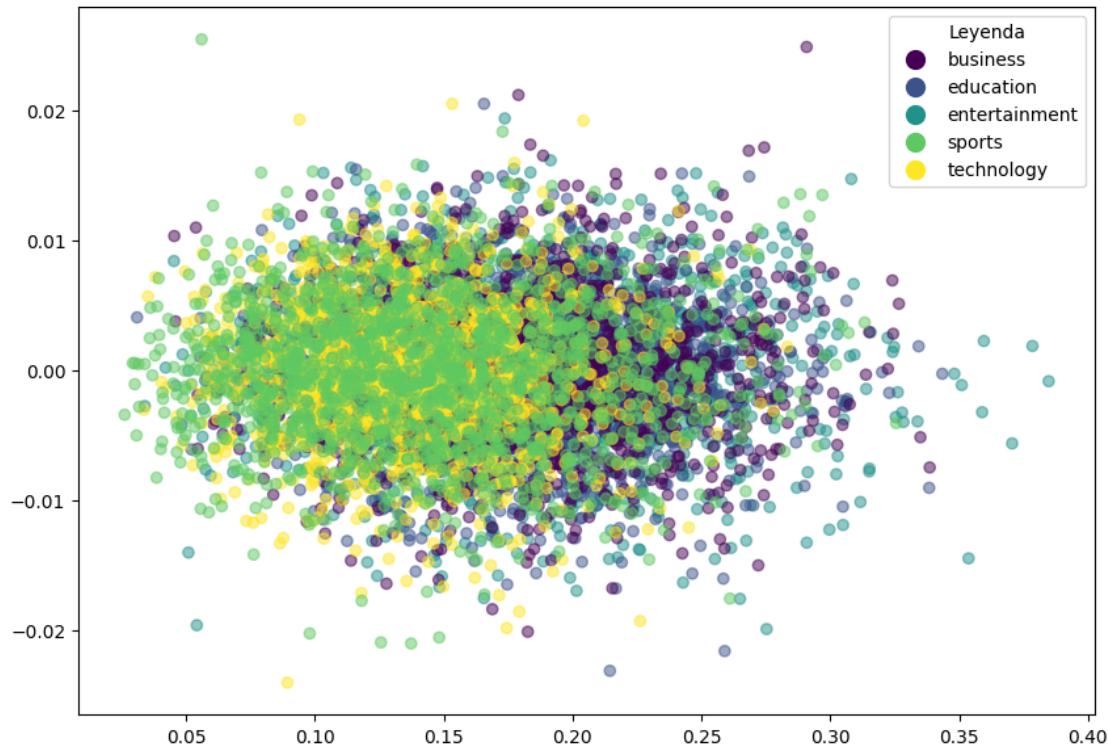
```
[51]: draw_components(unique_labels, tfidf_df, Y_train, 4000)
```

```
(8000, 4000)
0.902929715798673
```



```
[52]: draw_components(unique_labels, tfidf_df, Y_train, 5000)
```

```
(8000, 5000)
0.945929726721801
```



```
[68]: draw_components(unique_labels, tfidf_df, Y_train, 6000)
```

```
(8000, 6000)
0.9741974647433137
```

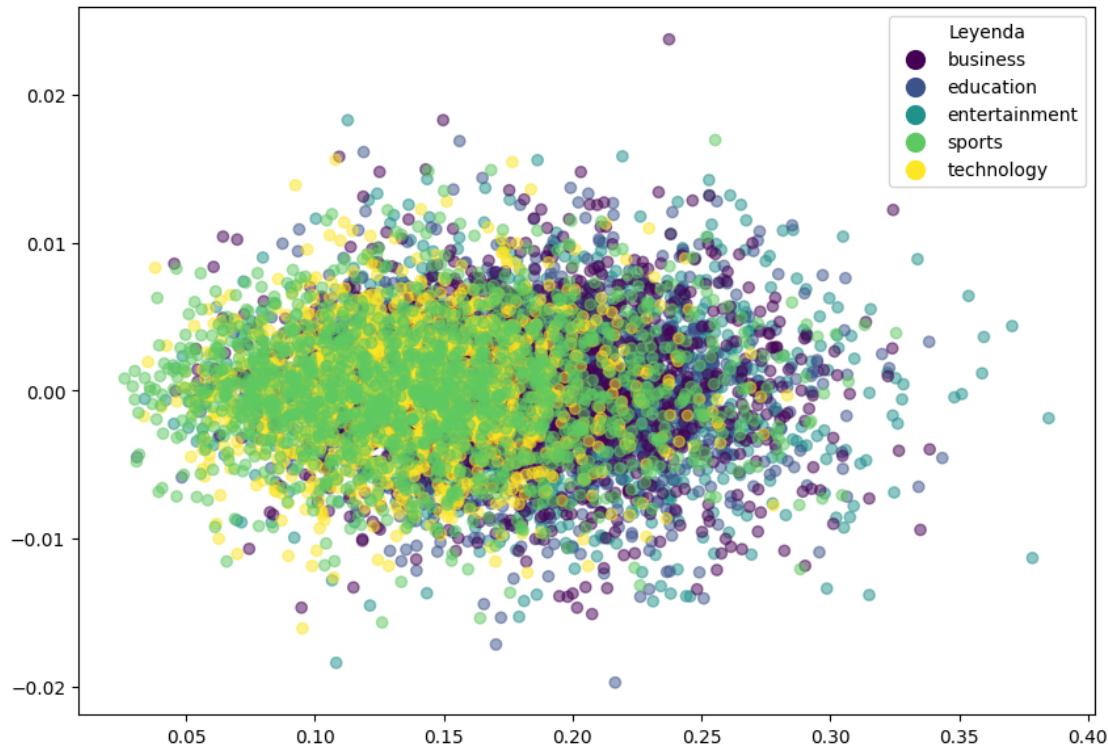
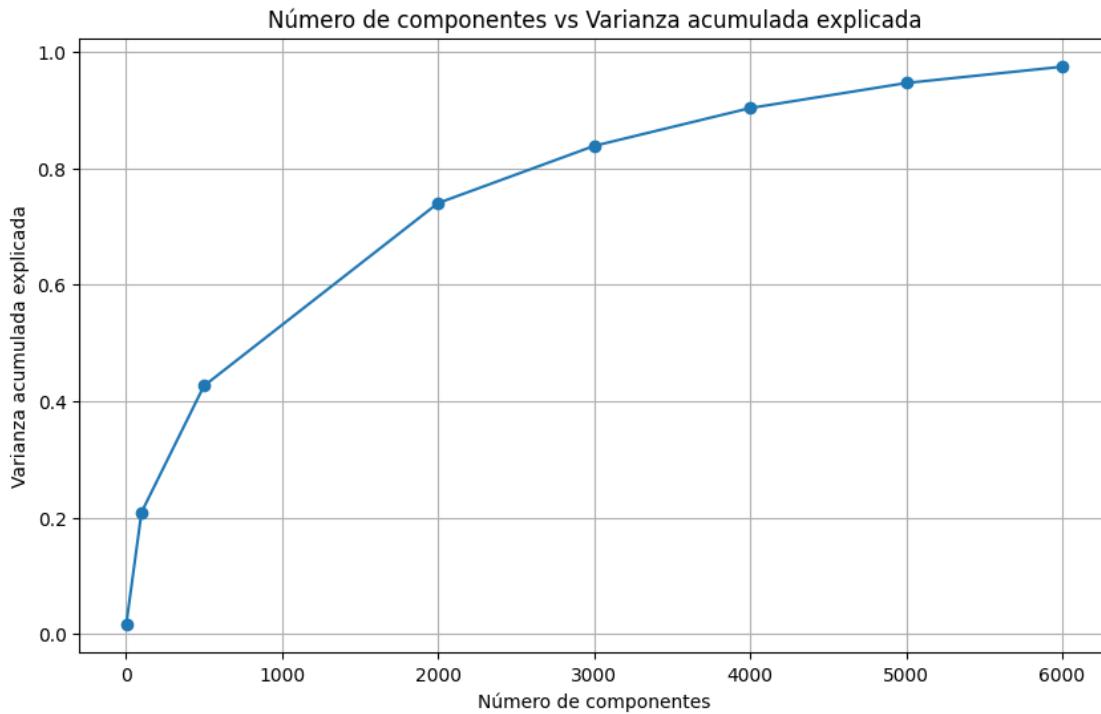


Tabla comparativa de la exploración de componentes principales

Explained Variance	Number of components
0.0161	2
0.209	100
0.426	500
0.7398	2000
0.838	3000
0.903	4000
0.946	5000
0.974	6000

```
[70]: x = [2,100,500,2000,3000,4000,5000, 6000]
y = [0.0161, 0.209, 0.426, 0.7398, 0.838, 0.903, 0.946, 0.974]
plt.figure(figsize=(10,6))
plt.plot(x, y, marker='o')
plt.xlabel('Número de componentes')
plt.ylabel('Varianza acumulada explicada')
plt.title('Número de componentes vs Varianza acumulada explicada')
plt.grid(True)
plt.show()
```



Se observa que con un mayor número de componentes hay una mayor varianza explicada. En este caso una varianza explicada del 0.94 es suficiente para explicar la mayoría de la información. Lo anterior debido a que al elegir un umbral del 94% se está asegurando que la mayoría de la información relevante se conserve en los componentes seleccionados, lo cual puede ayudar a garantizar que el modelo capture la estructura subyacente de los datos. Por lo que se elige tener una varianza del 0.94 con un número de componentes de 5000. Esto a su vez, reduce significativamente la dimensionalidad de los datos, lo que simplifica el modelo y hace que sea más fácil de interpretar, así como mejora la eficiencia computacional en comparación a si se escogen 6000 componentes.

Se define una clase para preparar los datos:

```
[101]: class TextPreprocessing():
    def __init__(self, stopwords=stopwords.words('english')):
        self.stopwords = stopwords
        self.pca = None
        self.tfidf_vect = None

    def remove_non_ascii(self, words):
        """Remove non-ASCII characters from list of tokenized words"""
        new_words = []
        for word in words:
            new_word = unicodedata.normalize('NFKD', word).encode('ascii', 'ignore').decode('utf-8', 'ignore')
            new_words.append(new_word)
        return new_words
```

```

def to_lowercase(self, words):
    """Convert all characters to lowercase from list of tokenized words"""
    new_words = []
    for word in words:
        new_word = word.lower()
        new_words.append(new_word)
    return new_words

def remove_punctuation(self, words):
    """Remove punctuation from list of tokenized words"""
    new_words = []
    for word in words:
        new_word = re.sub(r'[^w\s]', '', word)
        if new_word != '':
            new_words.append(new_word)
    return new_words

def replace_numbers(self, words):
    """Replace all interger occurrences in list of tokenized words with textual representation"""
    p = inflect.engine()
    new_words = []
    for word in words:
        if word.isdigit():
            new_word = p.number_to_words(word)
            new_words.append(new_word)
        else:
            new_words.append(word)
    return new_words

def remove_stopwords(self, words):
    """Remove stop words from list of tokenized words"""
    new_words = []
    for word in words:
        if word not in self.stopwords:
            new_words.append(word)
    return new_words

def stem_words(self, words):
    """Stem words in list of tokenized words"""
    stemmer = SnowballStemmer('spanish')
    stems = []
    for word in words:
        stem = stemmer.stem(word)
        stems.append(stem)
    return stems

```

```

def lemmatize_verbs(self, words):
    """Lemmatize verbs in list of tokenized words"""
    lemmatizer = WordNetLemmatizer()
    lemmas = []
    for word in words:
        lemma = lemmatizer.lemmatize(word, pos='v')
        lemmas.append(lemma)
    return lemmas

def stem_and_lemmatize(self, words):
    words = self.stem_words(words)
    words = self.lemmatize_verbs(words)
    return words

def preproccesing(self, words):
    words = self.to_lowercase(words)
    words = self.replace_numbers(words)
    words = self.remove_punctuation(words)
    words = self.remove_non_ascii(words)
    words = self.remove_stopwords(words)
    return words

def transform_train(self, X, n_components):
    X_train_new = pd.Series(X)
    X_train_new = X_train_new.apply(contractions.fix)
    X_train_new = X_train_new.apply(word_tokenize)
    X_train_new = X_train_new.apply(lambda x: self.preproccesing(x))
    #X_train_new = X_train_new.apply(lambda x: self.stem_and_lemmatize(x))
    X_train_new = X_train_new.apply(lambda x: self.stem_words(x))
    X_train_new = X_train_new.apply(lambda x: ' '.join(map(str, x)))
    self.tfidf_vect = TfidfVectorizer()
    X_tfidf = self.tfidf_vect.fit_transform(X_train_new)
    self.pca = TruncatedSVD(n_components)
    X_train_pca = self.pca.fit_transform(X_tfidf)
    return X_train_pca

def transform_test(self, X, n_components):
    X_test_new = pd.Series(X)
    X_test_new = X_test_new.apply(contractions.fix)
    X_test_new = X_test_new.apply(word_tokenize)
    X_test_new = X_test_new.apply(lambda x: self.preproccesing(x))
    #X_train_new = X_train_new.apply(lambda x: self.stem_and_lemmatize(x))
    X_test_new = X_test_new.apply(lambda x: self.stem_words(x))
    X_test_new = X_test_new.apply(lambda x: ' '.join(map(str, x)))
    X_tfidf = self.tfidf_vect.transform(X_test_new)
    X_test_pca = self.pca.transform(X_tfidf)

```

```
    return X_test_pca
```

```
[102]: pipeline = TextPreprocessing()
```

```
[103]: X_train_p = pipeline.transform_train(X_train, 5000)
print(f"El tamaño es: {X_train_p.shape}")
X_train_p
```

El tamaño es: (8000, 5000)

```
[103]: array([[ 7.26843905e-02, -7.39314058e-02,  2.07455779e-02, ...,
   -5.42930880e-03, -5.74008487e-03,  9.71890375e-03],
   [ 2.15273963e-01, -2.48027810e-01,  5.92943399e-02, ...,
   -5.52320161e-03,  4.90768423e-03, -4.31585830e-03],
   [ 1.75810585e-01, -9.48198861e-02,  4.50603119e-02, ...,
   -7.78025241e-03,  1.53078323e-03,  6.29265350e-03],
   ...,
   [ 1.55353659e-01, -5.66916740e-02, -3.61235092e-02, ...,
   2.50591523e-03, -2.02225171e-03, -8.09466202e-04],
   [ 5.12386126e-02, -1.46747521e-02, -1.20392566e-02, ...,
   -7.92149008e-04, -2.87622973e-03, -7.44831823e-05],
   [ 1.25058456e-01, -3.61265397e-02, -3.11409546e-02, ...,
   -1.16930123e-02, -1.06428496e-02, -4.13430106e-03]])
```

```
[104]: X_test_p = pipeline.transform_test(X_test, 5000)
print(f"El tamaño es: {X_test_p.shape}")
X_test_p
```

El tamaño es: (2000, 5000)

```
[104]: array([[ 2.96858919e-01, -2.12399520e-01,  1.87148730e-02, ...,
   1.04767224e-03,  3.81252977e-04,  7.21712974e-03],
   [ 1.28185073e-01, -6.59264962e-02, -1.73154613e-03, ...,
   -3.89268889e-03,  1.68675311e-03,  1.68509817e-03],
   [ 1.66823155e-01, -2.12022605e-02, -5.64368352e-03, ...,
   1.89739969e-03, -4.26128032e-03,  8.68838872e-03],
   ...,
   [ 9.88463172e-02, -2.52023737e-02, -3.47662685e-02, ...,
   4.06576363e-03, -3.70628069e-03,  8.65047293e-04],
   [ 5.94521234e-02, -6.02304744e-03, -1.94738607e-02, ...,
   -2.84723061e-03,  9.84046705e-06, -8.31764444e-03],
   [ 1.78927393e-01, -4.49383411e-02, -6.29069119e-02, ...,
   -4.87274434e-06,  8.46748489e-03, -1.19669703e-03]])
```

Se define la arquitectura de la red neuronal MLP. Se ven los tamaños de los datos de entrenamiento:

```
[105]: X_train_p.shape
```

```
[105]: (8000, 5000)
```

```
[106]: X_test_p.shape
```

```
[106]: (2000, 5000)
```

Número de clases:

```
[107]: len(unique_labels)
```

```
[107]: 5
```

5.6 Punto 3

```
[108]: model = Sequential(name="My_first_NN")
```

La capa de entrada:

```
[109]: model.add(Dense(128, activation='relu', input_shape=(X_train_p.shape[1],),  
                     name="Input_Layer"))  
model.summary()
```

```
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine  
Learning/Talleres/Taller3/env/lib/python3.11/site-  
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an  
`input_shape`/`input_dim` argument to a layer. When using Sequential models,  
prefer using an `Input(shape)` object as the first layer in the model instead.  
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
```

Model: "My_first_NN"

Layer (type)	Output Shape	Param #
Input_Layer (Dense)	(None, 128)	640,128

Total params: 640,128 (2.44 MB)

Trainable params: 640,128 (2.44 MB)

Non-trainable params: 0 (0.00 B)

Se define una capa oculta

```
[110]: model.add(Dense(64, activation='relu', name="Hidden_Layer"))
model.summary()
```

Model: "My_first_NN"

Layer (type)	Output Shape	Param #
Input_Layer (Dense)	(None, 128)	640,128
Hidden_Layer (Dense)	(None, 64)	8,256

Total params: 648,384 (2.47 MB)

Trainable params: 648,384 (2.47 MB)

Non-trainable params: 0 (0.00 B)

Capa de salida

```
[111]: model.add(Dense(len(unique_labels), activation="softmax", name='Output_Layer'))
model.summary()
```

Model: "My_first_NN"

Layer (type)	Output Shape	Param #
Input_Layer (Dense)	(None, 128)	640,128
Hidden_Layer (Dense)	(None, 64)	8,256
Output_Layer (Dense)	(None, 5)	325

Total params: 648,709 (2.47 MB)

Trainable params: 648,709 (2.47 MB)

Non-trainable params: 0 (0.00 B)

Ya con la arquitectura construida, se compila el modelo definiendo que función de pérdida, optimizador y métrica se va a utilizar para construir el modelo.

```
[112]: model.compile(optimizer='adam', loss='sparse_categorical_crossentropy',  
                     metrics=['accuracy'])  
model.summary()
```

Model: "My_first_NN"

Layer (type)	Output Shape	Param #
Input_Layer (Dense)	(None, 128)	640,128
Hidden_Layer (Dense)	(None, 64)	8,256
Output_Layer (Dense)	(None, 5)	325

Total params: 648,709 (2.47 MB)

Trainable params: 648,709 (2.47 MB)

Non-trainable params: 0 (0.00 B)

Se define un early stopping para verificar el aprendizaje y no esperar a entrenar todas las épocas si el modelo empieza a ver constancia en el aprendizaje

```
[113]: early_stopping = EarlyStopping(monitor='val_loss', patience=10, verbose=1,  
                                     restore_best_weights=True)
```

```
[114]: history = model.fit(X_train_p, Y_train, validation_split=0.2, epochs=100,  
                         batch_size=32, verbose=2, callbacks=[early_stopping])
```

```
Epoch 1/100  
200/200 - 1s - 5ms/step - accuracy: 0.7934 - loss: 0.7688 - val_accuracy:  
0.0000e+00 - val_loss: 6.3807  
Epoch 2/100  
200/200 - 1s - 3ms/step - accuracy: 0.9972 - loss: 0.0230 - val_accuracy:  
0.0000e+00 - val_loss: 7.3995  
Epoch 3/100  
200/200 - 1s - 3ms/step - accuracy: 0.9992 - loss: 0.0073 - val_accuracy:  
0.0000e+00 - val_loss: 7.7846  
Epoch 4/100  
200/200 - 1s - 3ms/step - accuracy: 0.9992 - loss: 0.0036 - val_accuracy:  
0.0000e+00 - val_loss: 8.1121
```

```

Epoch 5/100
200/200 - 1s - 3ms/step - accuracy: 0.9998 - loss: 0.0027 - val_accuracy:
0.0000e+00 - val_loss: 8.2311
Epoch 6/100
200/200 - 1s - 3ms/step - accuracy: 1.0000 - loss: 7.8149e-04 - val_accuracy:
0.0000e+00 - val_loss: 8.4242
Epoch 7/100
200/200 - 1s - 3ms/step - accuracy: 1.0000 - loss: 5.3663e-04 - val_accuracy:
0.0000e+00 - val_loss: 8.5851
Epoch 8/100
200/200 - 1s - 3ms/step - accuracy: 1.0000 - loss: 3.9260e-04 - val_accuracy:
0.0000e+00 - val_loss: 8.7210
Epoch 9/100
200/200 - 1s - 3ms/step - accuracy: 1.0000 - loss: 2.9839e-04 - val_accuracy:
0.0000e+00 - val_loss: 8.8457
Epoch 10/100
200/200 - 1s - 3ms/step - accuracy: 1.0000 - loss: 2.3342e-04 - val_accuracy:
0.0000e+00 - val_loss: 8.9558
Epoch 11/100
200/200 - 1s - 3ms/step - accuracy: 1.0000 - loss: 1.8642e-04 - val_accuracy:
0.0000e+00 - val_loss: 9.0569
Epoch 11: early stopping
Restoring model weights from the end of the best epoch: 1.

```

5.7 Punto 4

Gráfico del comportamiento de la pérdida en el entrenamiento:

```
[115]: plt.plot(history.history['loss'], label='Train')
plt.plot(history.history['val_loss'], label='Val')
plt.title('Training and validation loss')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.legend()
plt.show()
```

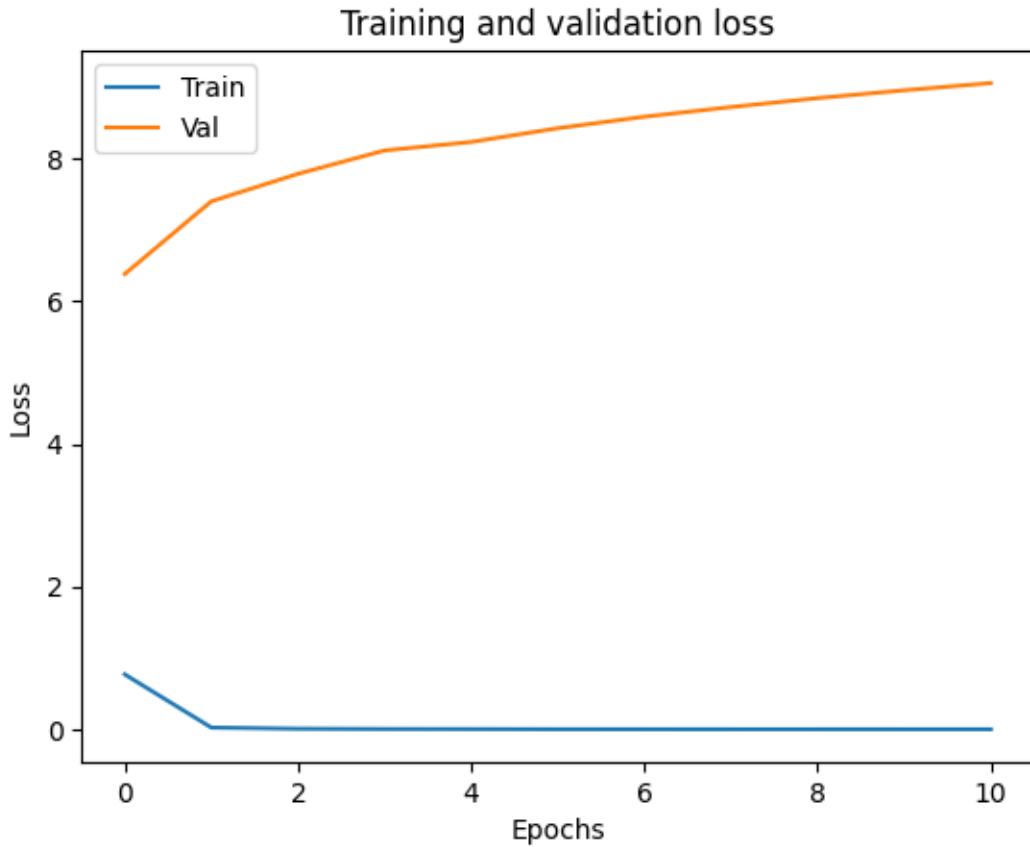
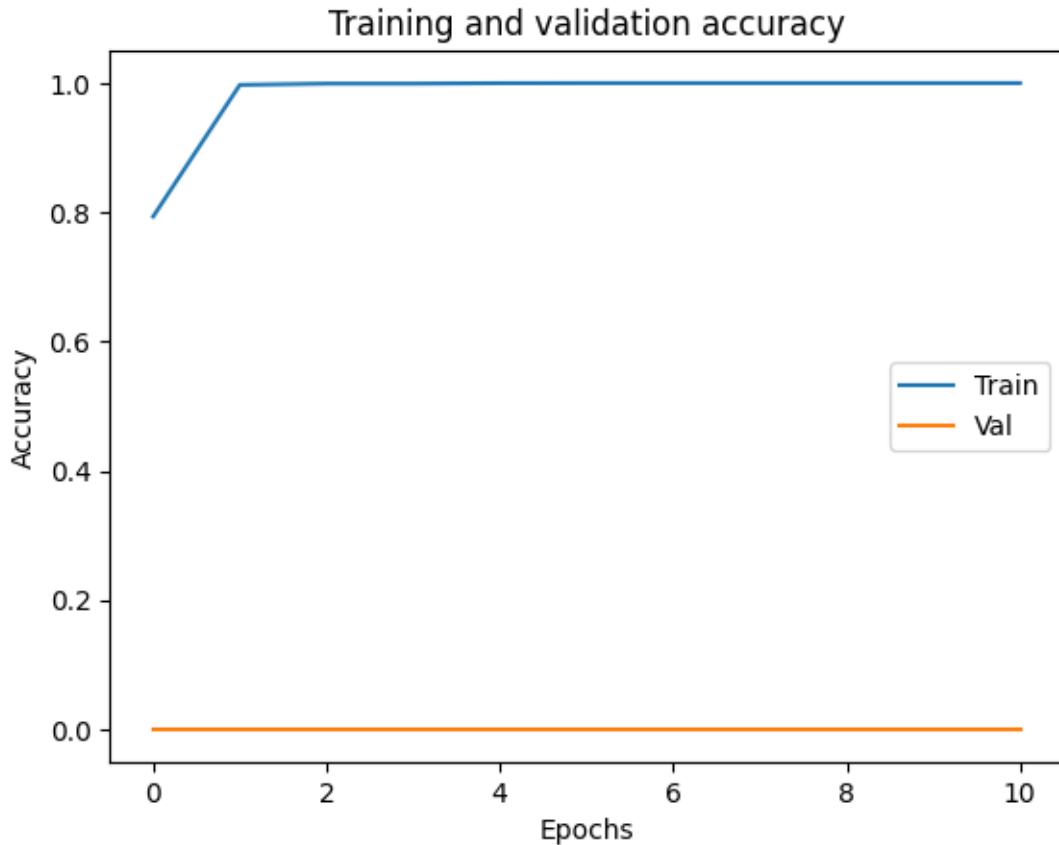


Gráfico del comportamiento de la métrica en el entrenamiento:

```
[116]: plt.plot(history.history['accuracy'], label='Train')
plt.plot(history.history['val_accuracy'], label='Val')
plt.title('Training and validation accuracy')
plt.xlabel('Epochs')
plt.ylabel('Accuracy')
plt.legend()
plt.show()
```



Se verifica la métrica con datos que no conoce la red (test)

```
[117]: model_accuracy = model.evaluate(X_test_p, Y_test)
print("Model Accuracy:", model_accuracy)
```

```
63/63          0s 1ms/step -
accuracy: 0.9555 - loss: 0.2521
Model Accuracy: [1.3602242469787598, 0.781000018119812]
```

Se visualiza la matriz de confusión del conjunto de datos train:

```
[118]: pred = model.predict(X_train_p, verbose=False)
predicted_classes = np.argmax(pred, axis=1)

print(classification_report(Y_train,predicted_classes,target_names=list(unique_labels)))

from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
cm = confusion_matrix(Y_train, predicted_classes)
disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=list(unique_labels))
```

```

fig, ax = plt.subplots(figsize=(8,6))
disp.plot(ax=ax)
plt.title("Matriz de confusión")
plt.tight_layout()
plt.show()

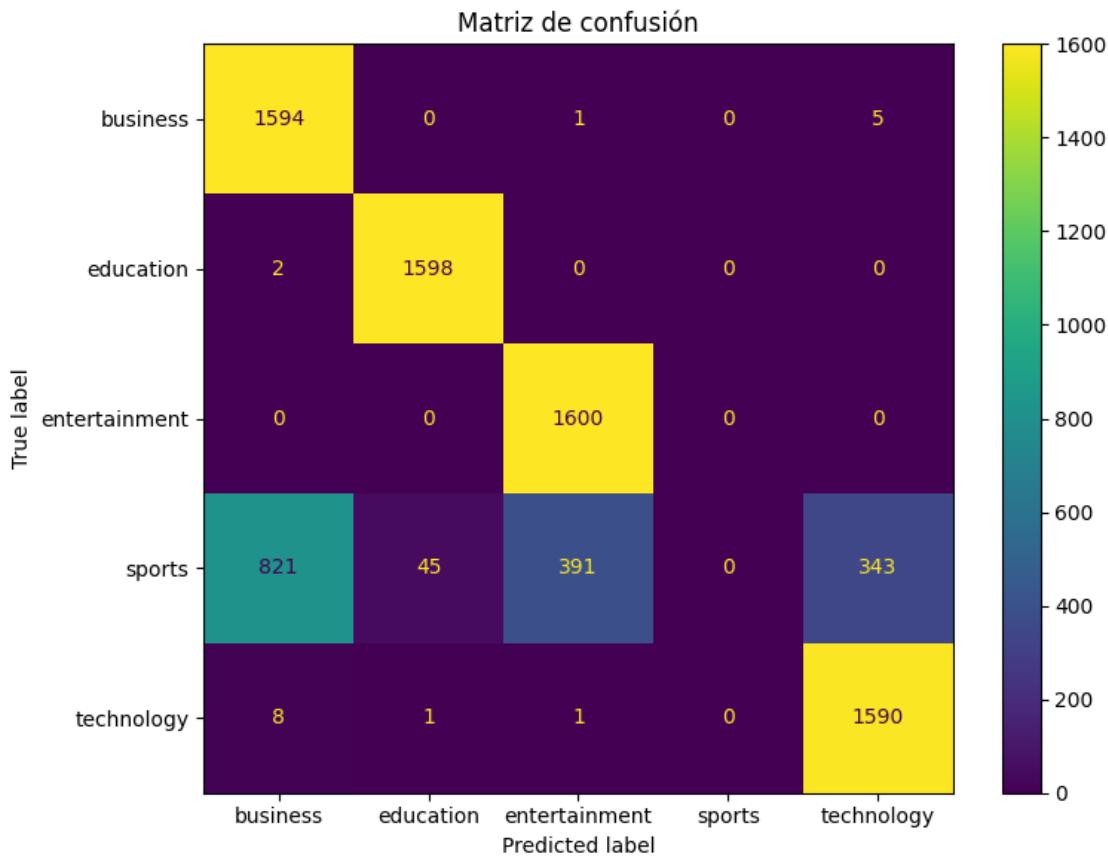
```

	precision	recall	f1-score	support
business	0.66	1.00	0.79	1600
education	0.97	1.00	0.99	1600
entertainment	0.80	1.00	0.89	1600
sports	0.00	0.00	0.00	1600
technology	0.82	0.99	0.90	1600
accuracy			0.80	8000
macro avg	0.65	0.80	0.71	8000
weighted avg	0.65	0.80	0.71	8000

```

/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning:
Precision is ill-defined and being set to 0.0 in labels with no predicted
samples. Use `zero_division` parameter to control this behavior.
    _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning:
Precision is ill-defined and being set to 0.0 in labels with no predicted
samples. Use `zero_division` parameter to control this behavior.
    _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning:
Precision is ill-defined and being set to 0.0 in labels with no predicted
samples. Use `zero_division` parameter to control this behavior.
    _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))

```



Se visualiza la matriz de confusión del conjunto de datos test:

```
[119]: pred_val = model.predict(X_test_p, verbose=False)
predicted_classes_val = np.argmax(pred_val, axis=1)

y_pred_val = predicted_classes_val

print(classification_report(Y_test,y_pred_val,target_names=list(unique_labels)))

cm_val = confusion_matrix(Y_test, y_pred_val)

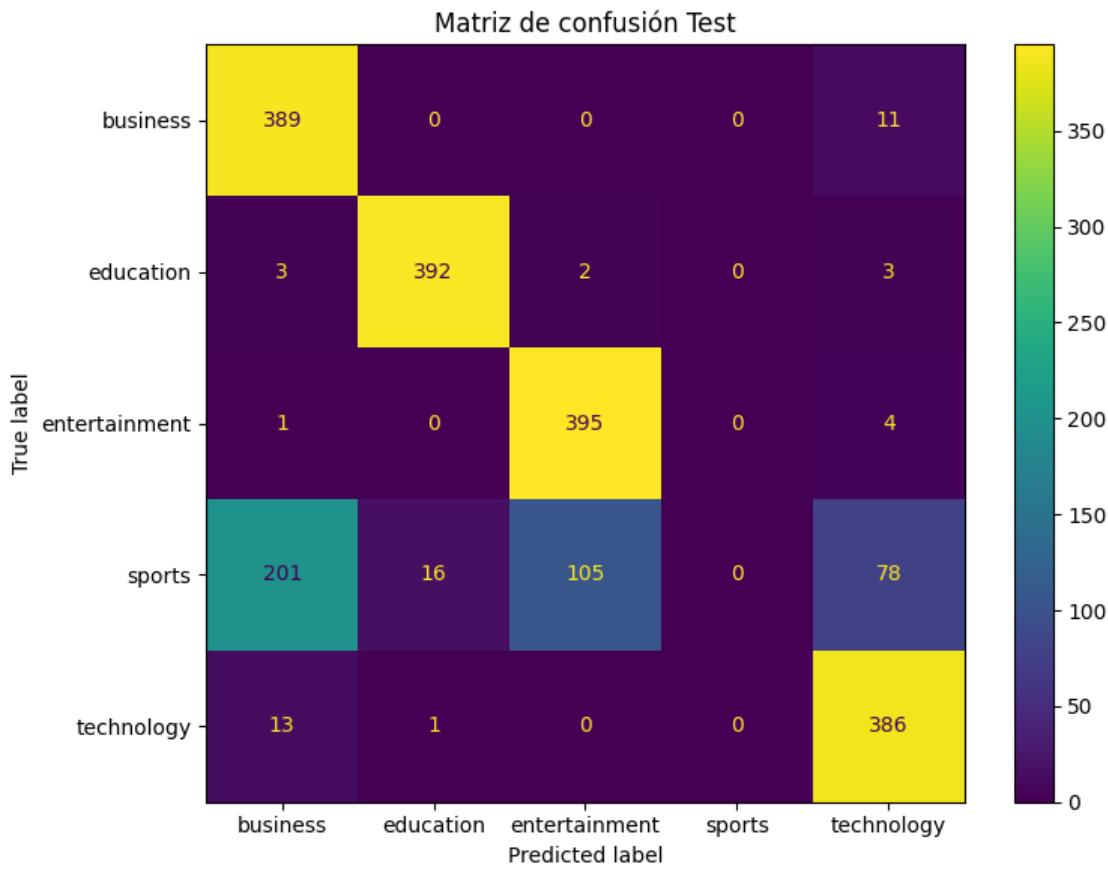
disp_val = ConfusionMatrixDisplay(confusion_matrix=cm_val, display_labels=list(unique_labels))

fig, ax = plt.subplots(figsize=(8,6))
disp_val.plot(ax=ax)

plt.title("Matriz de confusión Test")
plt.tight_layout()
plt.show()
```

	precision	recall	f1-score	support
business	0.64	0.97	0.77	400
education	0.96	0.98	0.97	400
entertainment	0.79	0.99	0.88	400
sports	0.00	0.00	0.00	400
technology	0.80	0.96	0.88	400
accuracy			0.78	2000
macro avg	0.64	0.78	0.70	2000
weighted avg	0.64	0.78	0.70	2000

```
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning:
Precision is ill-defined and being set to 0.0 in labels with no predicted
samples. Use `zero_division` parameter to control this behavior.
    _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning:
Precision is ill-defined and being set to 0.0 in labels with no predicted
samples. Use `zero_division` parameter to control this behavior.
    _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning:
Precision is ill-defined and being set to 0.0 in labels with no predicted
samples. Use `zero_division` parameter to control this behavior.
    _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```



En general, en el modelo baseline, se observan resultados prometedores para el entrenamiento. Con resultado promedio de un accuracy del 0.8. En este caso, se observan buenos resultados particularmente para las clases de educación, entretenimiento, tecnología y negocio. Por otro lado, se observa que se obtiene una baja precisión con la clase de deportes. Con respecto a los resultados de test se observa un accuracy de 0.78 por lo que se concluye que el modelo logró generalizar de manera adecuada lo cual lo ayuda a predecir de manera correcta frente a datos nuevos. Igualmente hacia adelante, uno de los posibles pasos a futuro sería realizar aumentación de la información y realizar modelos de interpretabilidad como SHAP para saber qué palabras está confundiendo entre clases y mejorar aún más las métricas.

Esto último, se sugiere dado que hay palabras que se repiten entre clases como se puede ver en el bag of words (Word Cloud) de la parte de limpieza de datos como por ejemplo: said.

Finalmente, como se puede ver en las gráficas el loss está aumentando en vez de bajar, por lo que una mejor función de regularización loss_function puede ayudar a mejorar los resultados a futuro también.

5.8 Punto 5

5.8.1 Búsqueda de hiperparámetros

Los hiperparámetros que se decidieron explorar fueron: optimizer, activation y batch_size. A continuación, se define cada uno de los parámetros:

- Optimizer: El optimizador es el algoritmo usado para actualizar los pesos de la red neuronal durante el entrenamiento. Algunos optimizadores comunes son SGD (Stochastic Gradient Descent), Adam, RMSProp, Adagrad, entre otros. La elección del optimizador puede tener un impacto significativo en la velocidad de entrenamiento y en la convergencia del modelo a mejores resultados.
- Activation: La función de activación determina la salida de cada neurona en la red. Algunas funciones de activación comunes son ReLU (Rectified Linear Unit), Sigmoid, Tanh, entre otras. La función de activación elegida puede afectar la capacidad de aprendizaje, por lo que se decidió seleccionar este parámetro.
- Batch Size: El batch size se refiere al número de muestras que se pasan a través de la red neuronal antes de que se actualicen los pesos. Un batch size más pequeño puede resultar en una actualización más frecuente de los pesos, pero puede requerir más pasos de entrenamiento. Se espera que para un batch size más grande, el modelo converja más rápido.

```
[120]: from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Conv1D, GlobalMaxPooling1D
from sklearn.model_selection import GridSearchCV
from scikeras.wrappers import KerasClassifier

def model_to_optimize(optimizer, activation):
    model = Sequential(name="NN_Hyperparameter_Tuning")

    model.add(Dense(128, activation=activation, input_shape=(X_train_p.
    ↪shape[1],), name="Input_Layer"))
    model.add(Dense(64, activation=activation, name="Hidden_Layer"))
    model.add(Dense(len(unique_labels), activation='softmax', ↪
    ↪name='Output_Layer'))
    model.compile(loss='sparse_categorical_crossentropy', optimizer=optimizer, ↪
    ↪metrics=['accuracy'])

    return model

params = {
    "model__optimizer": ["rmsprop", "adam", "sgd"],
    "model__activation": ["leaky_relu", "relu", "sigmoid"],
    "batch_size": [10, 20, 30],
}

model = KerasClassifier(build_fn=model_to_optimize,
                       epochs=20,
                       verbose=False)
```

```

search = GridSearchCV(estimator=model, param_grid=params,
                      cv=3, verbose=1, scoring="accuracy")
search_result = search.fit(X_train_p, Y_train)

/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.
    X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)

Fitting 3 folds for each of 27 candidates, totalling 81 fits

/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.
    X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.
    X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

```

```

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/scikeras/wrappers.py:925: UserWarning: ``build_fn`` will be renamed to
``model`` in a future release, at which point use of ``build_fn`` will raise an
Error instead.

X, y = self._initialize(X, y)
/Users/mariacatalinaibanezpineres/Desktop/MAESTRIA/2024-10/Machine
Learning/Talleres/Taller3/env/lib/python3.11/site-
packages/keras/src/layers/core/dense.py:86: UserWarning: Do not pass an
`input_shape`/`input_dim` argument to a layer. When using Sequential models,
prefer using an `Input(shape)` object as the first layer in the model instead.
    super().__init__(activity_regularizer=activity_regularizer, **kwargs)

```

[121]: `search_result.best_params_`

[121]: `{'batch_size': 10,
'model_activation': 'sigmoid',
'model_optimizer': 'rmsprop'}`

[122]: `search_result.best_score_`

[122]: 0.9827499994608702

[123]: `best_estimator = search_result.best_estimator_`

[124]: `test_result = best_estimator.predict(X_test_p)`

```
[125]: from sklearn.metrics import accuracy_score  
  
accuracy = accuracy_score(Y_test, test_result)  
print("Accuracy:", accuracy)
```

Accuracy: 0.9825

5.8.2 Análisis de los resultados

Luego de realizar el grid_search, se observó que el mejor optimizador fue RMSProp el cual introduce un coeficiente de atenuación. El optimizador RMSProp resuelve el problema de que el optimizador AdaGrad finaliza el proceso de optimización demasiado pronto; ya que, se usa el concepto de ‘ventana’ para considerar solo los gradientes más recientes.

Por otro lado, la función de activación que obtuvo mejor resultado fue la sigmoide. Esto se puede deber a que esta función transforma los valores obtenidos en un rango de 0 a 1, los valores altos tiendiendo a uno (1) y los bajos a cero (0). Esta además suaviza la salida de la red y es de fácil derivación, lo cual permite obtener valores para todos los puntos de la función.

Finalmente, el mejor batch size fue de 10. Esto se puede deber a que se toman pasos menores para llegar al mínimo global, lo cual puede ser más preciso y eficiente.

Después de hacer la búsqueda de hiperparámetros, se observó una mejora considerable en el accuracy. Esta aumentó de 0.78 a 0.98. Por lo que, se podría decir que hacer una búsqueda profunda de hiperparámetros sí influye y es esencial para mejores resultados.